ACCEPT

COOPERATION

Project description

Concept and objectives

Background and Need

As we move towards an ever more networked world, we are getting used to the idea that information is easily available. However, in this new paradigm something fundamental has happened: for thousands of years information has been created by the few and consumed by the many. Suddenly the power of the internet and its democratising effect has meant that anyone can in principle create information and make it available to anyone in the world with internet access. Yet the language barrier remains: however accessible information is, it is still only available to those who speak the language it is written in.

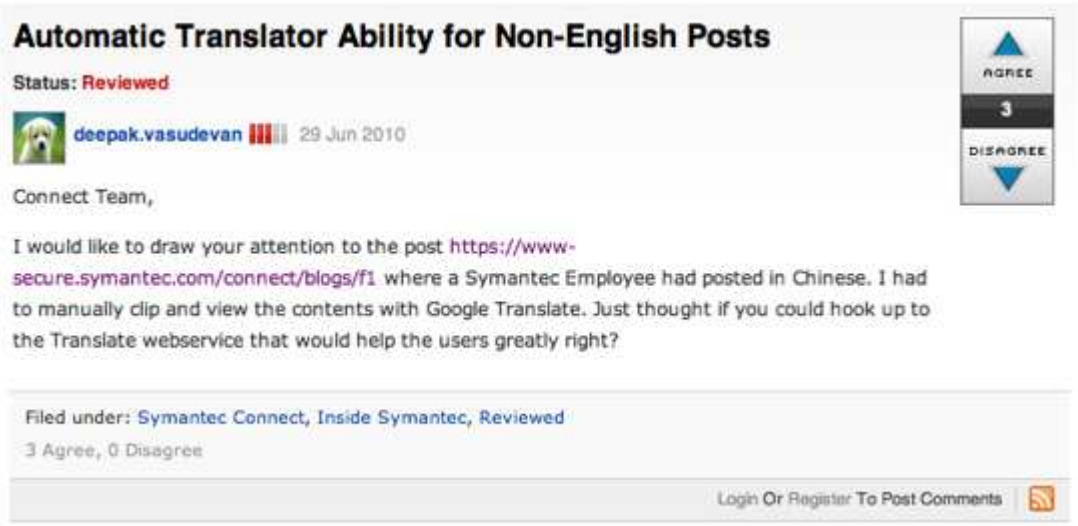
There are many definitions of "Web 2.0", but they all share the key concept of greater reliance on "the community", that is, the recognition of the fundamental shift from one-way communication between organisations (both commercial and non-profit) and their "users" to a community model, where the users are just as likely to help each other. There are many benefits to this paradigm: firstly it leverages a much larger pool of shared knowledge which users build up. Secondly, it is by its very nature very focused on actual user requirements, rather than those planned by the organisations. But perhaps the most significant benefit of user-generated content (UGC) for users is that it is created in the users' own languages. Many organisations struggle to translate all their content (typically they manage less than half), but users will naturally fill this gap in their own language.

In the case of companies like Symantec the community of users is organising itself and creating content around Symantec's products and services. The users create their own content in forums, answering each other's questions, and giving reviews of new features, and generally sharing experiences. Some of these forums are hosted and moderated by Symantec themselves, but there are also third-party sites creating useful content. A recent article claimed that 25% of all product searches now land on UGC. It is becoming a critical part of the user's experience. Furthermore, since the content is created by peers, users tend to trust this content and be more engaged with it - they are often willing to help each other just for the respect it can win them in the community.

The importance of UGC as a resource for customer engagement has been recognised by companies' efforts to encourage users to contribute content, and, especially, to reward users who deliver high-quality content. Forum members can earn stars or points in many forums, and in some cases even get rewarded with free software or other benefits. Why? Because companies are recognising that if their customers are happy, then they will be happier with their products, even if they had to help themselves. The content created by users is being co-opted as part of the user experience and as part of the support strategy of many major companies.

But of course this brings new problems, the biggest being the language barrier. When content is created by Chinese users it is only available to the Chinese-speaking community, unless it gets translated. Worse still, other language communities of users will not even be aware that the information exists - it will be created again for every community. For this reason, organisations are now addressing the challenge of translating useful user-generated content. Traditional manual or computer-aided translations are not capable of meeting this need, since they are too slow and expensive. The only viable solution is to use machine translation to provide translation on-demand.

This post sums up the user need:



However, machine translation of UGC is fraught with problems, even with Statistical Machine Translation (SMT) systems which are thought to be more robust. The key issue with UGC is the quality of the content. There are a number of reasons for this; very few of the users are professional writers and many of them are writing in a language which is not their mother tongue. To make matters more complicated, community forums typically have a rather informal tone, where poor spelling and punctuation, insider jargon, and strong sentiments are quite normal. Here's a typical example taken from a user post on Symantec's Norton forum:

*"I'm **of** on a weeks holiday next week so **i** may not be able to post till **i** get **back.i'll** have a look on my PC to see if **i** can find the long erase."*

But the new content creation paradigm is not just something which is commercially relevant. As Kofi Annan, the former General Secretary said; "Knowledge is power; Information is liberating". For many non-profit organisations (NGOs) the ability to deliver information is a crucial part of their mission. Basic health, nutritional or educational material can help save lives and help people help themselves. In crisis situations such as the earthquake in Haiti, medical care and basic food provisions were only part of the solution to helping people get their lives back. Information on how to use medicines and simple procedures to reduce the risk of diseases like malaria and cholera has to be made available to as many people as possible *in their own language*. Often the information consumers are individuals a long way from central distribution centres where English, French or other "major" languages are spoken - even the need for translation is often hard to communicate to those creating content.

Right now much of this life-giving information is created by the experts in their own language, and only a small percentage of it can be translated. Machine Translation (MT) is usually regarded as a risky strategy without editing, given the critical nature of the information. Using the power of the new community paradigm, Lexcelera provides free translations to NGOs through Traducteurs sans frontières (TSF), a network of volunteer translators. However there is always much more to be translated than there are volunteers available. MT would significantly help get more information to those who need it since, as with the Symantec product forums, the immediacy of the content presents a challenge in traditional translation workflows. However, here too there are significant challenges for MT: firstly the content often contains technical terminology which clearly needs to be correctly handled. The content creators are not professional writers, but doctors, nurses, engineers and teachers - who are often working under pressure in crisis situations. The benefits of light pre-editing for this kind of content are clear, but the pre-editing process needs to be as non-disruptive as possible.

Similarly post-editing is often needed for this kind of content since poor quality could be life-threatening. However traditional approaches to post-editing are not particularly helpful in this context, since the need for bilingual expertise is a major bottleneck. Experts who could potentially correct machine generated content in areas such as healthcare or engineering are not necessarily bilingual, and translators with that critical subject-matter expertise and knowledge of two languages are generally in shorter supply. This is as true of European languages as it is of African languages such as Swahili. To meet the pressing need to leverage MT effectively for this kind of content, we need to develop post-editing strategies which don't rely on source language knowledge - thus significantly increasing the pool of candidates for editing.

Several groups, both in commercial companies and research organisations have attempted to address these issues by fine-tuning MT systems to better deal with this kind of content. But as the history of MT has shown time and again, the most effective way to improve results is to improve the incoming content, the source text. The idea of improving content quality before translation has evolved beyond simple ideas of "controlled language" towards more targeted improvements which specifically increase the effectiveness of content as well as its translatability. Many organisations, including Symantec, have shown how fixing a relatively small number of linguistic phenomena can have a dramatically positive effect on MT results. The potential benefits of applying these principles of translation-friendly writing to UGC are clear - reliable MT would effectively remove the language barrier and give users access to all relevant UGC. As well as increasing user satisfaction, it would also help make companies more competitive in a multilingual environment.

However a significant challenge remains: how can this idea be applied to user-generated content, where the users are not employed to create content, as professional writers are - how can this kind of content be made translatable without demanding considerable editing effort from informal users?

The first major goal of this project is therefore to develop a user-friendly ("minimally intrusive") strategy for pre-editing User-Generated Content for Statistical Machine Translation. This involves identifying and evaluating the *minimal critical linguistic phenomena* for improving translatability and implementing these phenomena as rules in the acrolinx IQ system - the market leading software for language quality assurance, which Symantec has already deployed extensively for traditional

technical content. The rules will then be deployed directly in the content creation environments of the users.

This concept of “pre-editing lite” will be developed from the experiences of Symantec and others. We will look at the hundreds of linguistic phenomena which might be tested for during “editing” and then scientifically evaluate the effect that their application has on MT performance (in this case SMT performance, since the consortium already has considerable experience in editing for rule-based MT). Many of these phenomena have already been implemented in some form in the acrolinx IQ rules system, but have never been customised for this particular use-case. By carefully doing an independent and scientific cost-benefit analysis (cost = “time required to edit to the rule” vs. benefit = “improved translation performance”), we can arrive at the optimal set of rules. Furthermore, we can rank additional rules according to their benefits, so that users who want to spend more time editing can get the most useful rules first. The same principles can be applied to the evaluation of post-editing rules.

Analogous to the requirements of Symantec and their communities of users, the TSF community of NGOs is creating information, mainly medical, educational and nutritional, for use by people in areas of need with often very diverse linguistic backgrounds and language skills. Enabling effective MT for this kind of information which is often created by subject-matter experts rather than professional writers will significantly increase the reach of this information - and help TSF better achieve its mission of saving lives by delivering critical information in the right language at the right time.

Pre-editing in this context will also include the checking of critical terminology issues which directly affect translation accuracy. This is available as a standard feature of the acrolinx IQ system, but some research is needed to determine which terminology issues may lead to MT problems - always facing the compromises needed to avoid over-loading the content creators with too much pre-editing work.

For content which is not simply translated on demand, but is of enough significant value to justify editing, the MT output will often be post-edited to make it more comprehensible and more useful for the target audience. Currently this work is typically carried out by professional translators with expert knowledge of both source and target language as well as some subject matter expertise. Due to the specialised skills required for the task, post-editing still represents a significant bottleneck for the use of MT for anything other than gisting purposes. The state of the art with MT is such that we are still some years away from being able to deliver high-quality results without post-editing. In the case of TSF content, one of the main challenges with post-editing is to find enough human resources to deal with the output. Clearly MT systems can create almost infinite amounts of raw translated content, but traditional post-editing still requires bilingual expertise to process this output. If post-editing could be done by human resources with some subject matter expertise but just knowledge of the target language, this would significantly increase the amount of content which could be post-edited - thus further increasing access to that content and further lowering the language barrier.

The second major goal is to address the post-editing bottleneck. The project will work to identify the key factors which can make it possible for the post-editing process to be carried out by people with no knowledge of the source language - this is known as **monolingual post-editing**. Furthermore we will develop linguistic software to support the monolingual post-editing process.

There are basically two MT post-editing tasks:

- Re-ranking: Statistical MT systems give ranked results. Rules that focus on typical errors of these systems can be used to re-rank these results, such that the best translation is more likely to be at the top of the list of suggested translations.
- Reformulation: The rules can be used to reformulate erroneous sentences, containing for example spelling, grammar, or terminology errors.

While re-ranking can be done automatically, most of reformulation will have to be done by human editors. Post-editing is one of the most-unpopular tasks for professional translators. This is due to the fact that professional translators have a clear opinion about the translation task and high expectations about quality of the translation outcome. They are also typically paid less for post-editing than for translation, and many see it as a degradation of their work. There is an emerging trend towards specialised post-editors, and companies that focus on offering post-editing services, but the need for post-editing to be done by translators represents a bottleneck in the vision of delivering post-edited content as broadly as possible.

Post-editing MT output is a tedious task, if:

- training data of SMT is of low quality
- input text to MT is of low quality
- output of the translation is expected to be of very high quality, or the expectations on translation quality are not clearly defined
- there is no tool available that detects errors and provides suggestions for corrections

Post-editing in ACCEPT will be both a motivating and concise task. Training data of SMT will be corrected using acrolinx IQ, such that the MT output can be expected to be of higher quality than MT output that was based on unedited input data. The post-editing effort will be significantly reduced if we can succeed in encouraging pre-editing: The input to MT will be free from spelling and grammar errors, follow the MT translatability rules and have consistent terminology. For example, there is no need to convert passive to active (better suited for technical documentation), if this was already done in the process of pre-editing and training data editing.

The expectations on output will be clear: Output shall be comprehensible and correct, but not necessarily stylistically perfect. The post-editing effort will be reduced to the following tasks:

- spelling (upper and lower case)
- grammar (word order, agreement)
- punctuation, hyphenation
- check for untranslated terminology (and provide feedback to the MT system)
- terminology inconsistencies such as correct training data, translation memories (TM)
- use of wrong word in context (feedback to terminology checking in pre-editing, bilingual terminology and training data)
- formatting

Most important, the post-editor will be supported by automatic error detection with an adapted set of error detection rules, reduced to the error types that are necessary to make the text comprehensible and correct. The behavior of the editor will be logged, such that error markings that are often ignored can be turned off and new rules can be added on basis of free-text editions. By this method, the post-processing rule set will be adapted to what the editors actually need to detect and will constantly strive to reduce the editing burden to a minimum depending on the quality requirements of the translation task. As with pre-editing, it is a major deliverable of the project to collect insights and develop technology around a careful scientific evaluation of the cost-benefits of the editing process, to ensure minimum effort for maximum effect.

As mentioned above the goal of the project is to have a system than can continue to improve over time with the goal of minimising the pre- and post-editing effort. One of the most promising areas of research is to look at how post-edits can be fed back into the SMT system, so that it can produce the edited result directly. The goal is not necessarily to completely obviate the need for post-editing, but simply to automate it wherever possible.

The third major goal is to improve learning and develop feedback loops to improve SMT results for community data. SMT systems can learn well from large amounts of similar content in a single domain. Some work is needed to improve their performance in areas where parallel data is relatively sparse and the content is broader in domain - or may be completely out-of-domain. In addition to working on improving learning strategies for SMT in this kind of scenario, we will also develop methods for learning from human post-editing.

The ACCEPT project will create sophisticated components for supporting NLP applications, including machine-translation. One key part of this approach will be to try to “know more” about the content. By analyzing the content we will try to determine if there are predictable improvements we can make to the translation process depending on what we know about the content. For instance if we know that the content contains a procedure with steps or a set of questions for troubleshooting an issue, this information might be used to focus the efforts of both human and machine participants in the translation process - special editing rules can be activated and the SMT system can offer different rankings to better match this kind of content.

Another significant dimension of content is sentiment. If we can determine that content is heavily charged with positive or negative sentiment, this information would be extremely useful in ensuring that the translation correctly conveyed those sentiments. The system could perhaps even be set up to check that sentiment polarity (whether an utterance was positive or negative) was maintained from source to target. The effect of sentiment on MT quality is an area which has been largely ignored up to now, since MT has been mainly used for published content which is not heavily loaded with sentiment.

The fourth major goal is to improve reliability of forum content translation using Text Analytics. We will develop strategies and implement linguistic software to deliver useful insights into content. This component will include developing existing text classification algorithms for this purpose as well as rules for identifying sentiment and other relevant phenomena in informal forum content.

Users

To establish the proof-of-concept as clearly as possible the project will take a multilingual approach. The main focus will be on content created in English and French, which will also act as target languages for the project, in addition to German and Japanese. The project will look into the feasibility of supporting so-called sparse data languages like Swahili or Lithuanian in the future, but for which the appropriate linguistic resources do not yet exist. To ensure that these solutions are as generic as possible it will be tested in two slightly different scenarios. Firstly, for content in a typical commercial product forum relating to Symantec network security products. In the second scenario, we will address the needs of non-profit organizations through Traducteurs sans frontières, where content is created, translated and disseminated by NGOs and volunteers. TSF was created by Lexcelera to address the challenge of removing the language barrier for non-profit organisations, to help them better serve their "users" - the people they are trying to help. Despite the superficial differences between these two scenarios, the technological requirements are fundamentally the same. It is also expected that there will be considerable opportunities for the project users (Symantec and Lexcelera) to learn from each other and create innovation through this synergy.

The communities do not have to be built from scratch; TSF already has a platform and hundreds of users. Symantec has tens of thousands of active forum contributors. Our expectation is that we will play a significant role in driving growth of these user communities further.

The success of the project will to some extent be defined by the ability to show usage of the system by the user communities addressed in the project. We will take user acceptance extremely seriously during our evaluations, as the main focus of the project is to provide an “acceptable” level of support. Finding the balance between ease-of-use and the benefits of higher quality will be a matter needing considerable attention. The users in the project (Symantec and Lexcelera) will further develop programs in their communities to raise awareness of quality issues in community information.

The dissemination effort around the project will also address the issue of increasing the awareness amongst users that taking quality seriously can significantly increase the reach that their content can have. To put it simply, applying the ACCEPT system to their content will increase the number of people who can benefit from that information.