Procesamiento del Lenguaje Natural, Revista nº 53, septiembre de 2014, pp 193-196 recibido 15-04-14 revisado 02-07-14 aceptado 02-07-14

Integration of a Machine Translation System into the Editorial Process Flow of a Daily Newspaper

Integración de un sistema de traducción automática en el entorno de redacción de un periódico

Juan Alberto Alonso Martín, Anna Civil Serra
Lucy Software Ibérica SL
c/ Copèrnic 44, 1r 08021-Barcelona
juan.alonso, anna.civil@lucysoftware.com

Resumen: El artículo describe el proceso de integración del traductor automático de Lucy Software en el entorno de redacción de *La Vanguardia*, donde se utiliza diariamente como herramienta auxiliar para publicar una edición bilingüe del diario en catalán y en castellano. Este proceso de integración incluye adaptaciones técnicas y lingüísticas, y un proceso final de post-edición.
Palabras clave: traducción automática, post-edición, integración de la traducción automática en procesos productivos

Abstract: This paper describes the integration process of Lucy Software's machine translation system into the editorial process flow of *La Vanguardia* newspaper, where it is used on a daily basis as a help-tool in order to produce bilingual editions of the daily newspaper in Catalan and Spanish. The integration process includes both technical and linguistic adaptations, as well as a final post-edition process.
Keywords: machine translation, post-edition, integration of machine translation into productive processes.

1 Introduction

Established in 1881, *La Vanguardia* is the leading daily newspaper in Catalonia and the fourth best-selling in Spain, with a daily circulation of over 200.000 copies. It is widely recognized as a quality newspaper

In 2010 *La Vanguardia* decided to prepare a parallel edition in Catalan, which was officially launched on May 3rd 2011. In order to be able to do this parallel edition in Catalan, they decided to use post-edited Machine Translation (MT), and after surveying possible candidates, they finally chose Lucy Software's MT system.

This paper describes different aspects on the integration of the Lucy LT machine translation system into the editorial process flow of *La Vanguardia*. This integration involved both linguistic and IT aspects. You can find more details on this integration in Vidal and Camps (2012).

2 The Lucy LT MT System

Lucy LT is a rule-based machine translation system which is the ultimate successor of the old METAL MT system. Lucy LT is a transfer-based MT system with an island chart parser and three translation phases: analysis, transfer and generation. In each of these phases, and for each language-direction, computational grammars – one analysis grammar, one transfer grammar and one generation grammar -, and computational lexicons – source and target language monolingual lexicons and one source-to-target transfer lexicon – are used. Lucy LT runs on Windows workstations and has a number of APIs (e.g. Web Services) that allow integrating the system within external applications and

ISSN 1135-5948

workflows. For more details, please refer to Schwall and Thurmair (1997).

3 The Challenge

Whatever the final solution was, the following general requirements had to be met:

- One daily copy of *La Vanguardia* includes over 60.000 words, all of them to be translated, revised and post-edited.
- Both editions should be ready for printing every day at 23:30 the latest.
- The Catalan edition should comply with the linguistic requirements stated in the Style Guide of La Vanguardia.
- Even though most journalists at La Vanguardia write in Spanish, which was the base edition at the time, out of which the Catalan edition was to be created, at short/mid-term every journalist should be free to write in the language of his/her choice (Catalan or Spanish), so that, actually, after some time, there should be no base edition.
- Both the MT-system and the post-edition environment should be completely integrated into their editorial flow (both IT-integration and human team integration).

4 Possible Solutions

Given the task of making bilingual daily editions of a newspaper, three possible options could be considered:

4.1 The *MT-less* Option

This option would imply using no MT at all. This would imply:

- Duplicating the whole editorial human team or/and hire a team of N human translators to translate the entire newspaper content on time in order to keep both editions synchronized for publishing.
- Duplicating most of the IT infrastructure (Content Management System, etc.)

Given these factors, the question arises of whether it would be feasible to produce bilingual editions of a newspaper this way because of dramatic increase of costs and very tight time constraints.

This approach was therefore rejected.

4.2 The *full-MT* Option

This option would imply using only MT, without any post-editing phase. This means running all the contents of the (Spanish) base edition through an MT translation system and publishing the raw MT-translation of the original contents in the Catalan language edition.

It was immediately clear that this was not an option because, even for language-pairs for which the quality of MT is very high (as it is the case for Spanish-Catalan, where a quality higher than 95% can be achieved), the output mistakes would be unacceptable for publishing: proper nouns being translated, homographs, etc. Moreover, the Catalan style coming out from the MT system would not always sound "natural" to Catalan speakers.

This approach was also rejected.

4.3 The Sensible-MT Option

This option implied using a customized MT-system and a team of human post-editors. This option implied:

- Customizing the MT-system grammars and lexicons to the specific linguistic needs of La Vanguardia (style guide, corporate terminology, proper nouns, etc.).
- Integrating the MT-flow within the newspaper editorial flow (document and character formats, connection to a post-edition environment, feedback processing, etc.)
- Incorporating a post-edition environment to be used by a team of human post-editors into the editorial flow.

Here we have a compromise between the MT-use (time and effort saving) and the translation quality, so this was the approach that was finally chosen.

5 The Solution

The solution that was finally adopted by *La Vanguardia* implied the following general aspects:

5.1 Pre-launch Phase

There was a pre-launch ramp-up phase during which computational linguists from Lucy Software, post-edition experts, and part of the editorial team from *La Vanguardia* worked together for six months in order to

- Customize the MT-system to the linguistic requirements posed by *La Vanguardia* (as far as possible). This linguistic customization implied that over 20.000 lexical entries had to be added/changed in the MT-system lexicons and many grammar rules had to be adapted in the MT-system grammars, mainly for the Spanish→Catalan direction in a first phase.
- Integrate the MT-system into their IT editorial environment. This integration included:
    - The integration of our MT-system with *La Vanguardia*'s HERMES CMS.
    - Enabling Lucy Software's MT system to be able to handle *La Vanguardia*'s specific character format and XML tags.
    - Inclusion of markups in the MT-output specifically designed for post-editors
    - Configuring the MT-system installation so that translation performance could meet the expected translation load & peak requirements.
- Last, but not least, a team of around 15 persons were trained on post-editing the MT-output before publishing, and the corresponding shifts and work-flow for these post-editors was organized.

5.2 Post-launch Phase

In the post-launch phase, the lexicons and grammars of the MT-system continued to be adapted to the news that were translated every day in the Spanish→Catalan direction. Adaptation works also started for the Catalan→Spanish direction, also both in the lexicons and in the grammars of the system, in order to enable *La Vanguardia* journalists to write in the language of their choice (i.e., Spanish or Catalan).

This post-launch phase lasted for some six months right after the launch of the Catalan edition of the newspaper.

5.3 Maintenance Phase

The maintenance phase started right after the final of the post-launch phase and involves ongoing maintenance works, mainly in the computational grammars of both directions, Spanish→Catalan and Catalan→Spanish.

Previous to this, a training session was done with personnel of the newspaper's editorial team in order for them to get familiar with the lexicon coding tool of the Lucy MT system. Therefore, during this maintenance phase they are taking care of the system lexicons and Lucy Software is responsible of providing at least two annual updates of the computational grammars, where a number of reported errors have been fixed.

Beside the computational lexicons and grammars, the system has a so-called pre- and post-editing filters which allow the users to define strings that should not be translated (typically proper nouns). These filters are maintained by the staff of *La Vanguardia*, with the technical support of Lucy Software.

5.4 Examples of Linguistic Adaptations

Most of the MT lexicons adaptations that have been carried out for *La Vanguardia* correspond to

- Specialized lexicon entries on very specific domains:
    - Bullfight: albero/arena, morlaco/toro (bull)
    - Castellers (human towers): **cinc de vuit amb folre i manilles** (human tower of eight levels of five persons each), **pila de set**, etc.
- Proper noun lists, including lists of place names (villages, rivers, mountains, etc.), well-known person names (**Leo Messi**, **Rodríguez Zapatero**, etc.), etc.
- Latin words and expressions (**in dubio pro reo**, **tabula rasa,** etc.).
- New words (neologisms) or fashion words (Spanish/Catalan): **dron**/**dron** (drone), **bitcoin**/**bitcoin**, **autofoto/autofoto** (selfie), **crimeano**/**crimeà** (Crimean), **watsap**/**watsap** (a Whatsapp message))
- Words that appear often at *La Vanguardia*: **perroflauta/rastaflauta** (anti-system young person), **cantera azulgrana/planter blaugrana** (Barcelona F.C. team), **iniestazo/iniestada** (a score from Andrés Iniesta), **arena política** (political arena), **stajanovista/estakhanovista**.
- Idioms or colloquial language: **tartazo/cop de pastís** (pie hit), **hacer un corte de mangas/fer botifarra** (a rude gesture somehow similar to a *two-finger salute*), **cocinillas/cuinetes** (*kitchen wizard*, sometimes said in a derogatory sense).

6 Post-Edition

As already mentioned, the post-edition phase is a key factor in the final output quality of the Catalan edition. The post-editors typically work from 17:00 to 23:00. Their main goal is to revise and eventually correct mistakes that may appear in the MT output, and – also very important – give the *human flavor* to the MT Catalan output, whenever it is possible because of time constraints. The reason for this last point is that often, the MT output can be perfectly correct from a grammatical point of view but, still, sound a little *awkward or artificial* to a native speaker, in the sense that s/he would never use these words or construction to express this idea. The post-editor task is then to paraphrase the output sentence with a more *natural* wording. Again, because of time constraints, this is typically done in news headlines, where, in addition, more often than not puns can be used in the source language that are impossible to be correctly translated (or actually, localized) into the target language by an MT system.

7 Conclusions

The conclusions of this project can be summarized as follows: producing two parallel bilingual editions of a daily newspaper only seems to be feasible if the following three conditions are met:

- MT is used in the process,
- The MT-system is properly customized, adapted and integrated to the newspaper linguistic and IT requirements,
- There is a team of trained specialized human post-editors who correct MT mistakes and "give the human flavor" to the output.

References

Bernardi, U., Bocsak, A & Porsiel, J, 2005.: *Are we making ourselves clear? Terminology management and machine translation at Volkswagen.* Proceedings of the 10th EAMT conference "Practical applications of machine translation", 30-31 May 2005, Budapest*;* pages 41-49.

Schwall, U. & Thurmair G., 1997. *From METAL to T1: systems and components for machine translation applications*. Proceedings of the MT Summit VI. Machine Translation Past, Present, Future. Proceedings, 29 October – 1 November 1997, San Diego, California, USA; pages 180-190.

Vidal, B. & Camps, M., 2012: *Catalan Daily Goes Catalan* (www.localizationworld.com/lwparis2012/presentations/files/A4.pdf) presented at Localization World 2012, Paris.

Wolf, P., & Bernardi, U, 2013: *Hybrid domain adaptation for a rule based MT system.* Proceedings of the XIV Machine Translation Summit, Nice, September 2-6, 2013; ed. K.Sima'an, M.L.Forcada, D.Grasmick, H.Depraetere, A.Way; pages.321-328.