Lexicon-Coding Workflow at CLS Communication

Hans-Udo Stadler

CLS Communication AG
Elisabethenanlage 11, CH-4051 Basel
hans-udo.stadler@cls-communication.com

Abstract. Offering Machine Translation (MT) as a commercial service to end users means having to build up and regularly update large lexicons that can handle both general language and client specific terminology. This paper describes the processes developed at CLS Communication to cover the demand of users from several fields. Over a period of four years, processes have been set up and constantly refined in order to tackle the issue of improvement of MT output incorporating the resources at disposal. Different approaches have been tested, that have led to the implementation of three main coding streams. The coding of „unknown words", translation analysis and exports from terminology databases enable the MT team at CLS to react to changing external and internal needs. Accuracy and expenditure of time have been evaluated, as have the applications of automation and human skills.

1. Introduction

The initial idea for CLS Communication's MT service came from the clients themselves, that is from a Swiss Bank whose employees needed a solution for fast and inexpensive translations. They were looking for a way to translate mainly internal communications to and from several languages within a secure environment. CLS Communication had already provided them with human translations and other language services and thus had the competency and synergy potential to offer them the requested service.

2. Implementation

Building an MT system from scratch would have meant years of development apart from having to recruit the necessary experts from fields which did not belong to the company's core competencies. CLS Communication therefore decided to acquire a license to use and adapt an existing MT system. In 2000 and 2001 several commercial systems were evaluated with regard to translation quality and scalability, among other aspects (Maier and Clarke, 2001). Finally, the transfer system by Comprendium Lingua (formerly SailLabs) was chosen and installed in a server as well as a desktop version. One of the main advantages of this particular software was the lexicon development component called LexShop, which allowed CLS to not only add new complex terminology but also edit all entries, including the very sophisticated linguistic information needed for the translation software's rules. After all, specialized lexicons resulting in quality improvement are the main added value or, from a clients' perspective, the main reason to choose CLS Machine Translation.

3. Target Groups of CLS MT

Most of CLS' clients are companies from the finance, telecommunications and pharmaceutical sectors. MT is offered to them online so they can perform the translations themselves, or they can opt for our post-editing service, where the MT output is revised by a human translator who removes the most serious errors and selects the correct alternatives. Those translators can be regarded as internal clients for the MT team, since they proved to have their own wishes and needs with respect to our MT services (Hyland, 2003). Over the medium term, we are also planning to integrate CLS Machine Translation as a pre-translation tool in the regular human-translation workflow in order to increase productivity.

4. Impact and General Conditions

Given the different target audiences and scenarios as well as the diversity in domains and translation directions (mainly DE, EN and FR in all combinations), the lexicon-coding workflow has to be flexible and efficiently controllable. While building up a team of an average five FTE over the past four years, we experimented with different focuses and processes. This allows us to draw on a wealth of experience and adapt our procedures to the respective coding project. At the same time we strive to keep these processes as simple as possible, so that each of the team members can perform all of the tasks. One crucial point to bear in mind is that each of them works on a lexicon of their own.

For the time being, the Comprendium system does not support concurrent multi-user coding, so each person uses a local copy of the current lexicon and saves their changes to temporary "patch" files, which are collected and later cleaned into the "mother lexicon" to build a new version. It is only then that all PCs and servers are updated with the same new lexicon. This procedure, of course, has an impact on all our coding tasks.

5. Coding Streams

Currently there are three main coding streams, which we incorporate into our lexicon work on a regular basis:

- "unknown words"
- translation analysis, and
- exports from terminology databases.

Additional coding projects, i.e. imports of glossaries, for example, usually combine various processes of those streams. Figure 1 shows a simplified presentation of our workflow.

5.1. Unknown Words

With each translation that our clients run on the server installation, so called "unknown words" (UNKs), i.e. words not contained in the lexicon,

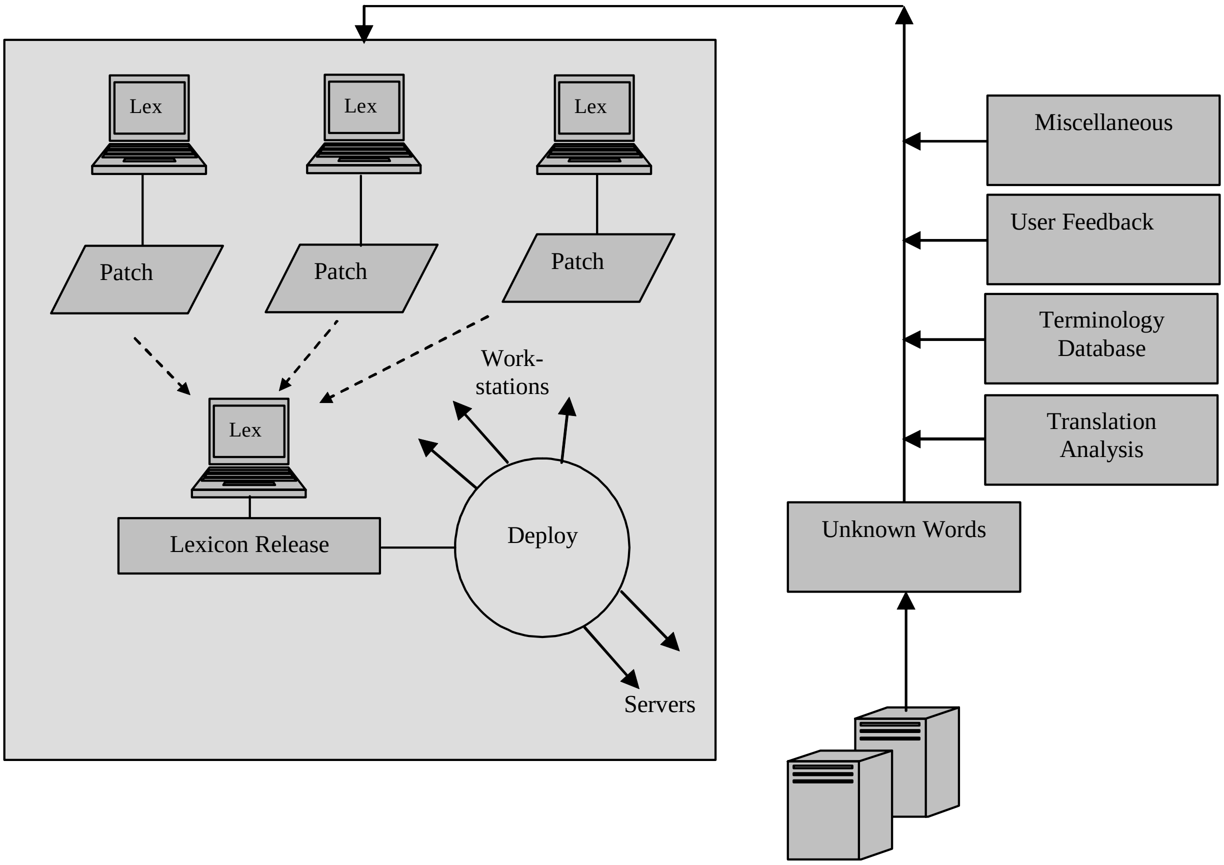

Figure 1. Coding Workflow (simplified)

are filtered and written into separate files. Since there is obviously a demand for translation of these words, we want to add the most frequent and relevant to the lexicon.

5.1.1. Preliminary Steps

Due to the configuration of our system, all UNK files per day and per server are generated within one folder, regardless of the translation direction. Handling thousands of individual files within the given folder structure proved to be very tedious and time consuming. All files, or at least the number of files that we decided to consider, had to be opened and then be sorted according to the translation direction. But then we still did not know how often the unknown words in question actually occurred. In order to automate the task of file handling, we had a script developed which collects all UNKs of a freely chosen period of time, and automatically sorts them by frequency.

The next steps have to be performed manually again, since they require certain linguistic as well as world knowledge: the list of unknown words has to be cleaned from the noise, mostly words that were misspelled or sent for the wrong translation direction, for example. If a common English word like „History" appears in the list of unknown words, it arouses one's suspicions. It might for example either have occurred within an English text which was translated from German into English instead of vice versa, or might have occurred as a foreign word within a German text.

The source language can be seen from the sentence which the word in question was found in. This particular sentence had been copied from the original text into the UNK file and kept for the list, so it can now be consulted to judge both the source language and the meaning within the specific context. The person in charge of cleaning the list can decide whether the unknown words are relevant for our lexicon or will likely not be sent for translation again, especially in the case of foreign words or peculiar compounds. Depending on the period over which the UNKs had been collected, there may still be thousands of „good" candidates left even after the noise has been deleted. Many of those had been sent for translation only once or twice and would not be worthwhile to add to the lexicon. We rather concentrate on a set of a few hundred of the most frequent UNKs, which usually have occurred at least five times. The rest, again, will be deleted from the list.

The finished list is then split into separate lists for each translation direction, which can be used for research of the target language equivalents. Depending on the language skills, research can be done by the same person or by a colleague.

5.1.2. Research of Unknown Words

For research we have a variety of sources, which we use depending on domains and clients. For very specific terminology, CLS Communication's vast translation memories and terminology databases are most productive, while more generic terms can be found in online or conventional dictionaries. The chosen translations for the unknown words are added to the list.

Optionally, a short quality assurance can follow, i.e. another colleague has a look at the finished list to make sure that the correct translations were chosen and no important second meaning of a term has been missed.

All in all, proficiency in both languages is necessary but research is different from human translation, where the translator should be native in the target language. For research we found that it is important to have very good knowledge of the source language because it is sometimes hard to understand what very specific expressions mean, and the person doing the research cannot be an expert in all fields. The quality assurance, on the other hand, is usually done by someone who is (near) native in the target language.

Before the new words are added to the lexicon, an administrative step is interposed. As described before, every team member has a personal copy of each lexicon and saves the changes to patch files which are used to build a new lexicon at regular intervals. Before these changes become effective on the online system, the same unknown words may be sent for translation by the users again and thus may occur in different UNK files over a certain period of time. In order to make sure that the same words are not researched and later coded by different team members, every list of researched words is copied into a central list before the coding has begun. By means of a macro, new research lists are matched with this central file and redundant occurrences deleted from the new list.

5.1.3. Coding of Unknown Words

Following the research, the new words can now be added to the lexicon. Again, this can either be done by the same person or by a colleague, depending on skills, personal preferences and workload. Coding usually consists of two main steps, semi-automatic import and manual editing. The entries are imported in the form of reformatted research lists by loading them to the transfer component of the lexicon. The respective monolingual source and target entries are defaulted by LexShop. Although the linguistic quality generated by the defaulting component is quite good, most entries have to be checked and corrected manually. When the respective team member has finished the corrections and saved the entries to patch files, those patch files can be loaded by a colleague for quality assurance and test translations. The revised entries are saved again to patch files, which are collected for the regular lexicon build.

5.2. Translation Analysis

Coding unknown words is fine for building up specific terminology but it does not guarantee improvement of overall translation quality. There are at least two other important aspects that have a grave impact on quality and can, at least partially, be controlled by lexicon coding:

- translation of words within a given context
- choice of equivalent according to domain.

The difference between “translation” and “equivalent” in this case means that first the system needs some guidance in recognizing what was meant in the source text in order to be able to choose between a set of correct possible equivalents. A verb, for instance, may have different meanings, of which one only applies if the verb is followed by a specific preposition and by a certain type of direct object. If either the preposition is not coded within the verb entry or the noun in question is not tagged as being of the respective type, the verb might be analysed in the wrong way. In the case of German it can become even more tricky, as many verbs are composed of a general verb and a prefix that can be separated, depending on the inflexion. Within a given sentence, the verb and the prefix might be analysed as two independent words if the subject is not coded with the correct grammatical information. So „translation of words within a given context” includes more than translation, namely the step of lexicon entry recognition. It is only after the successful recognition has been accomplished that the difficulty of choosing the right equivalent arises, provided that the desired equivalent is coded in the transfer component of the lexicon, at all. These two phenomena cannot be covered by coding unknown words but have to be addressed by translation analysis (TA). Basically, given source texts are machine translated and the MT output compared to the original, but there are certainly different ways to go about it. At CLS, we set up a specific stream of processes for that task.

5.2.1. Main Issues of TA

The first crucial step is finding appropriate texts for translation analysis. We usually choose from two categories: „live data”, i.e. source texts that were sent for Machine Translation by clients, and „parallel texts” from various sources. The live data is easy to get hold of, as it is temporarily stored on the MT server. Furthermore, it has the advantage of being relevant in that it contains the sort of texts our clients actually use MT for. Yet, the administration of this data is quite time-consuming, since the individual files have to be viewed and sorted, and only a fraction of the texts are suitable for translation analysis, in terms of length and style. Parallel texts, on the other hand, may not exactly be of the sort of texts that MT is usually recommended for but they readily provide us with the correct translations, which we can rely on, instead of having to look up single words in various sources.

Translation analysis, of course, reveals some unknown words, but we mainly concentrate on additional meanings of existing lexicon entries. Sometimes it is also sufficient to change the subject area for existing transfers, in order to improve the result. Above all, translation analysis allows us to identify collocations and phrases which can be added to the lexicon in the form of fixed expressions or flexional multiwords.

Compared to the UNK stream, there is clearly much more manual work involved in translation analysis. First of all, one has to deal with a variety of word categories while unknown words mainly concern nouns and can therefore be treated in list form. But different categories require different grammatical information, so lists would be cumbersome to prepare. Especially verbs are very difficult to handle, both with regard to monolingual and transfer entries. In most cases, they call for a lot of trial and error because one cannot foresee all the effects a change implies. Moreover, having entries defaulted by LexShop is of some help but comprehensive adjustments are indispensable. In short, automating translation analysis is hardly feasible while exhaustive testing is a must. Nevertheless it is worth the effort, because it definitely helps to improve overall translation quality significantly. Many common expressions are coded that do not only benefit users from single work scopes. In fact, general phrases seem to be more important for comprehensiveness than specific terminology. As the correctness of these phrases is easy to judge and the person doing the analysis already looks at the entries in detail, we usually dispense with an additional quality assurance.

5.2.2. TA and Post-Editing

As mentioned before, CLS Communication also offers a rapid post-editing of machine translated documents. That task is performed by translators who are native speakers of the target language. The result is not a human translation, in terms of quality, but sometimes it makes sense to use these documents, both source and target, for translation analysis as well. The raw MT output is compared to the original, while the post-edited version, which should of course be used with caution, can be a helpful reference.

In one particular post-editing project we were able to considerably improve the MT output in cooperation with the translators involved. Over a period of several months, the client sent one document per day for translation. The original texts were always written by the same person and very repetitive regarding content and style. Each day, the translator in charge sent the source and the target documents together with a few comments to the MT coding team. They told us what was most annoying about the raw MT output and sometimes came up with suggestions on how to solve it. The solutions they wished for were of course not always realisable but we could fix quite a few problems. Here too, it was rather a matter of adding fixed expressions and changing the subject area for some entries than adding new terminology. As a result, we could provide the translators with a much better MT output for the rest of this particular project and also assume that some of the changes constitute an improvement for all users.

Apart from the differences described, the processes for translation analysis are similar to those for coding of unknown words.

5.3. Exports from Terminology Databases

CLS Communication’s terminology team hosts terminology databases for several clients. The databases are usually concept-oriented, that is terms that have the same meaning are entered in one record. For example, there are several entries for the word „premium”, one with the meaning „consideration paid for a contract of insurance” and another in the sense of „reward”. Within each of the two entries there might be synonyms and several target terms for all languages recorded.

One database is especially used in-house as a reference for translators. It is extended and updated on an ongoing basis and currently contains just under 60,000 entries with translations for each term in German, English and French. To make use of this resource we once added all database entries then existing to the MT lexicons, and have since been regularly exporting the new entries into Excel files, which are prepared for import into our MT lexicons. However, since the original data was meant for other purposes than MT, we faced several obstacles in this undertaking. First, one has to decide what exactly to export and how to go about it. Most of the additional information contained within each entry is of no use for our MT lexicon, so it does not make sense to export everything. Many, though not all of the entries feature some useful grammatical information like word category or gender. Here, it has to be decided whether to keep the information where present or to do without. In the first case, the entries have to be treated separately from those not having that kind of information, which means an additional effort. Not using the information, on the other hand, means disregarding parts of the human knowledge that had already flown into the database.

Also, the terminology database includes domain tags but the structure is fundamentally different from the one we use in MT. The domain information can of course be exported and then replaced systematically but mapping the two structures to each other needs to be well thought through.

The main problem indeed is the structure of the entries themselves, i.e. that each source language term could have a different number of translations for the various target languages. And these translations do not really have to be equivalents but might differ in meaning or usage. This fact is perfectly convenient for translators who can view all synonyms at once and choose the appropriate. But for us it means that we have to tidy up the exported data, because there is no way to control the order of the target terms during the export. We tried to get the best out of the export by writing filters and export definitions but that is only where the next challenge starts. The exported data has to be brought into a format that is accepted by the MT lexicon tool. Moreover, corresponding source and target terms have to be sorted where the wrong „synonyms” were exported together. For example, in the database „premium” and „option premium” (in the sense of „price which the buyer of an option pays to the seller when the contract is closed”) are treated as synonyms and thus appear in the same entry. The German translations „Prämie” and „Optionsprämie” are given as target terms within this entry. „Optionsprämie” should not be coded as a transfer for „premium” in the MT lexicon because it would be too specific and might be wrong within a certain context. So if they had been exported together we have to make sure not to add them to the lexicon.

On the other hand, missing information, like word categories, has to be added, and systematically differing values have to be replaced. In short, much manual work has to be done before the files are ready for import. And since not everything can be completed in the files, there is some more work waiting after the import.

In conclusion, we can say that terminology databases provide us with useful and relevant multilingual terminology but they are far from being easy to use for automatic imports.

5.4. General Issues

For the entire workflow it is necessary to ensure that all persons involved have access to the respective data via a common network drive. Guidelines for file names have to be followed so that everyone can trace information on the contents and the current status of each file. The sequence of the single steps is controlled by means of various task lists containing the path to the files in question and the name of the person in charge of the current step.

Building a new lexicon by loading and releasing the completed patch files requires some time. All the more so, since it is impossible to avoid duplications between coding tasks of different team members. The deployment of the new lexicons to all workstations and the servers can in part be automated but for security reasons we refrain from full automation.

6. Conclusion

Development of the lexicon-coding workflow at CLS Communication has shown that balance and flexibility are of central importance. Each of the three main coding streams described has its own advantages. While unknown words reflect the obvious demand of current users, translation analysis reveals more general linguistic shortcomings. Multilingual Terminology databases, on the other hand, make research superfluous and thus save time.

CLS Communication’s MT team makes complementary use of all three coding streams and also re-uses a combination of processes of each stream for individual coding projects that do not fall into one of the three categories.

Occasional shifts in priorities and focuses proved to deliver best results with regard to translation quality.

7. Acknowledgements

The author would like to thank all members of the Machine Translation team at CLS Communication AG and Urs Hengartner of Canoo Engineering AG.

8. References

MAIER, E. & CLARKE, A. (2001). ‘Evaluation of Machine Translation Systems at CLS Corporate Language Services AG’. Proceedings of ‘MT Summit VIII’, 18-22 September 2001, Santiago de Compostela, Galicia, Spain; pp 223-229.

HYLAND, Catherine. (2003). ‘Testing Prompt: The Development of a Rapid Post-Editing Service at CLS Corporate Language Services AG’. Proceedings of ‘MT Summit IX’, 23-27 September 2003, New Orleans, USA; pp.189-193.