Streamlining Terminology Management in an RBMT Context

Horst Liebscher
text&form
Director of Technology and Innovation

horst_liebscher@textform.com

Thomas Senf
text&form
Managing Partner

thomas_senf@textform.com

Abstract

While the awareness for the importance of effective terminology management processes has grown significantly in the recent past, the starting point for terminology projects or terminology work is often less than ideal: In many cases, usable terminology assets do not exist at all, are of unknown origin or out of sync with the relevant content.

Efficiently establishing terminology requirements that incorporate existing assets and creating smart terminology management workflows are almost impossible without linguistic support. Content optimization tools or more generic linguistically based text analytics tools are not always available or do not find their way into the corporate budget because of their often prohibitive cost. However, if an inherent need for MT exists—a global corporate strategy can be reason enough to set up a company-wide MT portal for internal gist translation—two birds can potentially be eliminated with one stone:

While an RBMT system can be easily and elegantly integrated into the corporate terminology management workflow, the results that can be achieved by tapping into the inherent linguistic intelligence of an RBMT system beyond its perceived conventional purpose may provide additional justification for purchasing and maintaining such a system.

Our presentation will outline a practical approach towards this goal.

1 Introduction

Any RBMT system generally “knows” a minimum of two languages. It can analyze a source language text and use its knowledge of the transfer relationships between the two languages to generate a target language text based on sophisticated rules.

In theory, a perfect RBMT system only fails when faced with incomplete / incorrect sentences or unknown or ambiguous terminology. This suggests that the quality of the RBMT output can be improved by providing better sources, by coding new transfer relationships between source and target and by defining which of several alternative target terms should be used in a given context.

Essentially, RBMT systems “feed” on terminology. An open, user-friendly system should report on unknown terms, provide feedback on existing alternative translations for the same source term and indicate if an essentially unknown compound word was “translated” based on the individual components that it assumes to know. All this data reflects the terminology requirements from an MT perspective, which may not always correspond to the requirements that are relevant to a human translation process.

Experience with both conventional translation projects and MT scenarios indicates, however, that there can be significant overlap between the two sets of requirements. We will demonstrate how the features of an MT system can be used to establish an efficient terminology workflow and provide robust linguistic support for the required processes.

Sima’an, K., Forcada, M.L., Grasmick, D., Depraetere, H., Way, A. (eds.)
Proceedings of the XIV Machine Translation Summit (Nice, September 2–6, 2013), p. 391–396.
©2013 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.

2 The Terminology Value Chain

First, we need to outline and define the fundamental phases of the suggested terminology management process:

1. Term casting
2. Classification
3. Qualification
4. Entry compilation
5. Implementation in the target applications

In Step 1, a text analytics / term extraction tool identifies suitable term candidates, optionally syncing the suggestions with existing terminology assets.

The next step classifies these proposals into true candidates [C], subterms (such as abbreviations etc.) that need to be linked to other main terms [L], spelling errors [S] und rejected proposals [N].

The categories [C] and [L] are further qualified as preferred terms [P], forbidden terms [F] or allowed terms [A].

Finally, the terms are compiled and preprocessed as monolingual entries for the intended data structures.

These can then be used to populate the target system(s), which comprise at least a terminology component for the translation process, and potentially a web-based application for publishing purposes and/or a content optimization component. In addition, the terminology assets in the text analytics component need to be updated.

3 The Subset Delta: Friend or Foe?

Our intention is to significantly enhance this workflow using a rule-based machine translation component that adds a completely new aspect to the process.

While dedicated terminology harvesting components—whether statistical or linguistic—are designed to extract term candidates from a text corpus and match the new candidates to existing assets, rule-based MT systems consider themselves “smart” when analyzing text corpora.

In reality, this means that parsed terms found in the system’s own general vocabulary will not automatically be suggested as term candidates. The RBMT system assumes to already “know” the terms and does not doubt their validity regardless of the subject area. Other linguistically based text analytics tools will typically harvest and suggest such terms, even if they are found in their own respective general dictionary. An approach that exclusively relies on an RBMT sys-

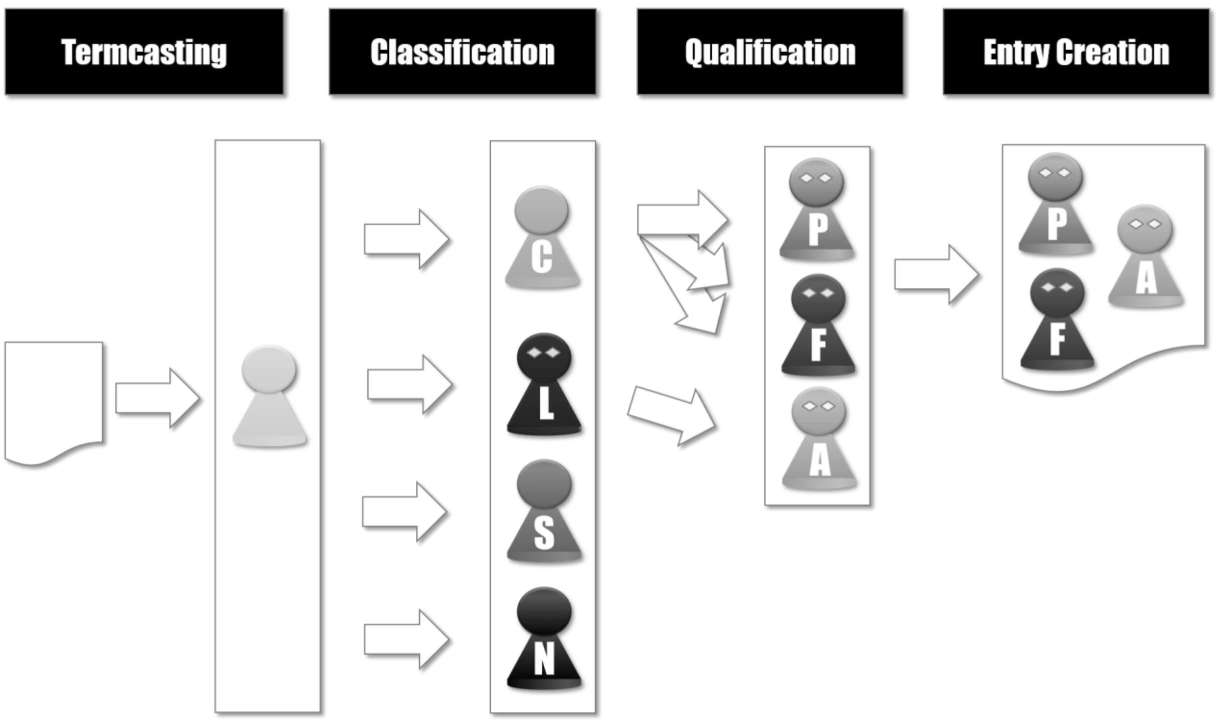

Fig. 1: The fundamental phases in the terminology value chain

tem for term harvesting will therefore initially extract fewer candidates.

The delta between the larger subset generated by the analytics tool and the smaller subset output by the RBMT system is unknown. Our proposed process aims to determine and quantify exactly what this delta is. To do so, we compare the output from the RBMT-based process with the results from two linguistically-based text analytics systems from a quantitative and qualitative perspective.

In addition, we will demonstrate how the very same RBMT system can be used to reduce the delta.

4 Part I: Capturing the Data

1. The suggested terminology process initially analyzes a test corpus in the RBMT system, which returns the following data:
   – unknown source terms,
   – alternative translations for the same source term,
   – ambiguous translations,
   – potentially unreliable ("synthesized") translations of compound words
   – any existing target language suggestions also provide additional information, e.g. the underlying domain (subject area)
   – the number of occurrences of the source term in the corpus can also be returned. All extracted term candidates are automatically reduced to their base form.
2. The same test corpus is run through a text analytics system [TA], which also provides a set of suggested term candidates including some linguistic metadata. The candidates in this set are also extracted in their base form.
3. All the data feeds into a centralized database system ["pentübrid Control – pCtr"] for systematical analysis and processing.
4. The pCtrl environment identifies the delta between the two sets of candidates, i.e. all terms that were extracted by the text analytics system but ignored by the RBMT system which assumes to know them already.
5. The delta is then passed from the pCtr environment to the RBMT system for translation; and the results feed back into the pCtr database.

5 Part II: Classifying the Candidates

In the next step, the term candidates are classified into three categories:

[A] Relevant terminology candidates for the translation process

[B] Candidates that are irrelevant to the translation process but may be necessary in an MT context

[C] Rejected proposals

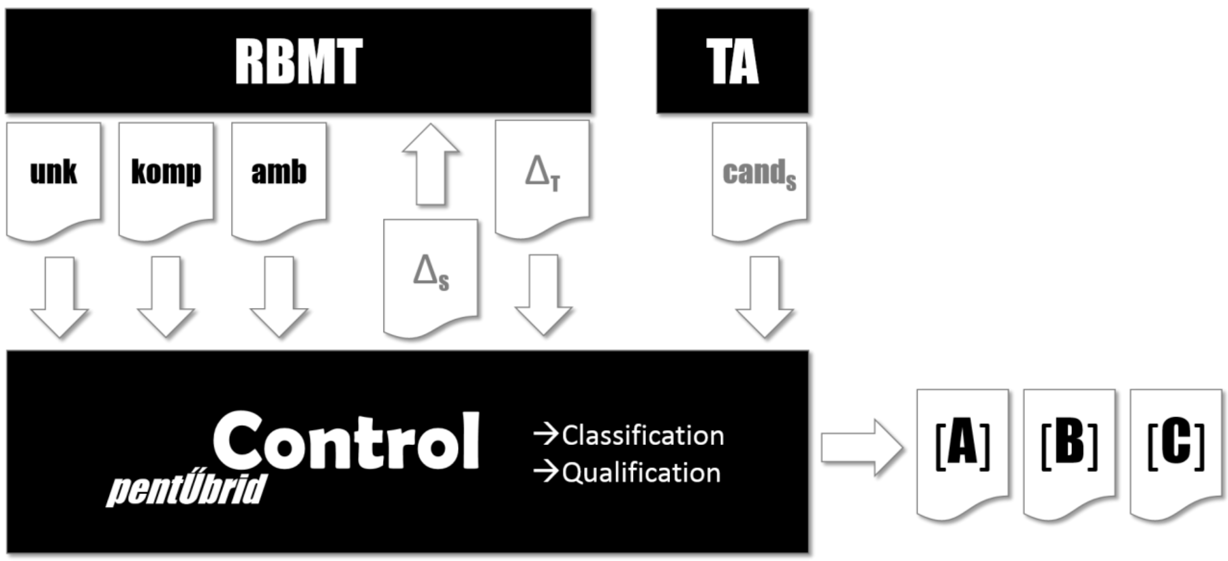

Fig. 2: Data flow around the pCtr system

While this in itself does not constitute a fundamental difference to the conventional process, all subsequent decisions can now be made on the basis of an automatic pre-classification that the pCtr system offers as an advantage over any existing constellation:

1. Common proposals from both systems (TA and RBMT): These are likely candidates for category [A].
2. Proposals specific to the text analytics component: These have already been translated by the MT system. If they are found suitable for category [A], a reviewable translation already exists. Otherwise, the candidates are moved to category [C].
3. Ambiguities or alternative translations, i.e. source terms for which the MT system holds multiple targets: These should always be specified for category [B] but may also be suitable candidates for category [A].
4. Compound words where the MT system only knows (or assumes to know) the individual components: These may be suitable candidates for either [A] or [B], and the suggested translation can be correct or at least helpful.
5. Proposals specific to the MT system: These likely candidates for category [B] can also be suitable for category [A], thereby adding substantial value to the overall process.

Once the classification process is complete, the various output formats relevant to the individual target systems (terminology component for the translation process, a web-based component for publishing purposes and the RBMT system itself) can be generated, and the assets can be synchronized.

As a major advantage over the conventional approach almost all candidates already come with one or several suggested translations in the target language of the RBMT system.

6 Benefits

First of all, the new process outlined in this document does not pose any disadvantages compared to a conventional linguistically-based terminology harvesting scenario. Any potentially useful candidates that the RBMT system initially misses are extracted by linguistically-based text analytics systems and subsequently pre-translated by the RBMT engine. This is a purely technical step that does not require an additional investment on the part of the user as it could be handled by external resources within a reasonable time frame.

In addition, the new process offers a variety of benefits:

– The automatic pre-classification in the system significantly accelerates the decision

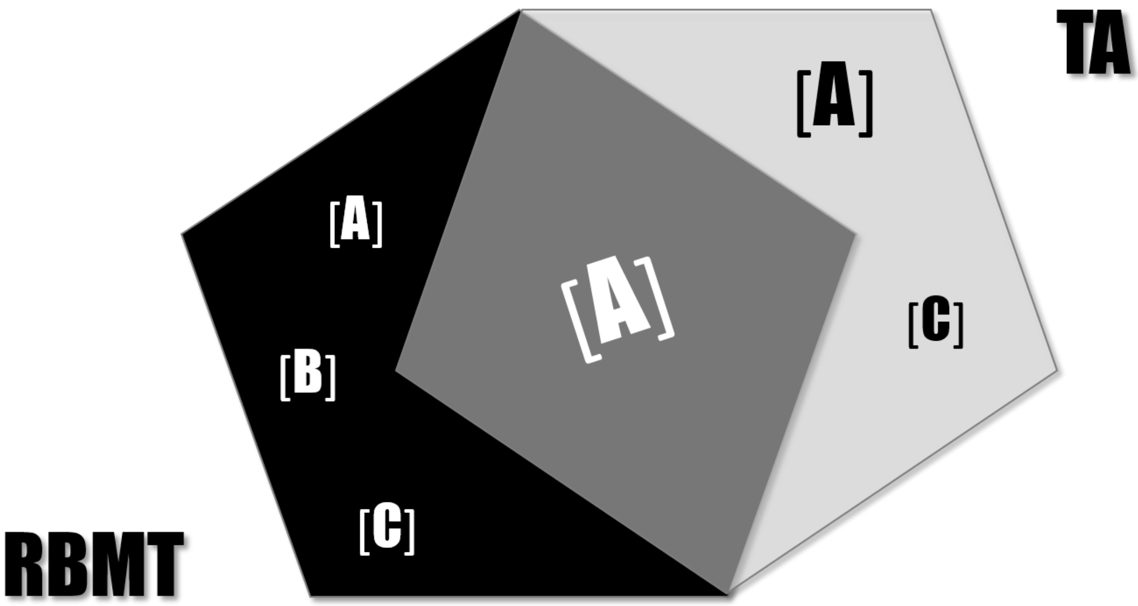

Fig. 3: Classifying and qualifying term candidates

processes required for classification and qualification.

– All proposals that were identified (extracted) by the text analytics component only come with a suggested translation from the RBMT system that then merely needs to be verified and validated.
– For a number of candidates, the RBMT system suggests alternative translations (possibly from various subject areas or client-specific domains) that can be validated or prioritized accordingly.
– The RBMT system itself suggests additional candidates that would not be identified in a conventional scenario. These are generally useful from an RBMT perspective but could also be suitable for the overall translation process. Otherwise, they can quickly be ignored or refused.

7 Test sets

Two discreet test sets were analyzed to illustrate the workflow. While the first set is a text of a more prosaic nature, test set 2 represents a section from a technical documentation.

A specific terminology domain was not previously set up for either test set in either the RBMT or the TA system.

8 Results: Quantitative aspects

The overlap between the candidate subsets harvested by the RBMT system and the TA components was not insignificant even for the more prosaic test set. The suggested process can be considered efficient and useful even with an intersection of "only" 20% compared to the conventional process.

For the technical documentation test set, the intersection was much higher, at around 50%.

The "intersection proposals," i.e. terms that were identified by both the RBMT system and the TA component(s) proved to be suitable term candidates without exception, which helps accelerate the classification process significantly.

9 Results: Qualitative aspects

At the time of writing this paper, the qualitative assessment of the results within the test sets had not been completed. The final results of the evaluation will be presented at the conference and will focus on the following questions in particular:

– How are rejected and accepted candidates distributed within the intersecting set of candidates between the two subsets, RBMT and TA?
– How does the ratio between the intersecting candidates and the overall number of proposed candidates change in light of this distribution?
– Is there a systematic pattern to the selection of good and bad term candidates followed by each of the systems using the available metadata, if necessary?
– Can general conclusions be drawn concerning the quality of the translation proposals generated?
– Can the choice of translation variants by the system be more systematically and reliably supported?

10 Conclusions

It can be concluded that the processes described here significantly increase speed and accuracy and offer highly practical advantages and efficient terminology interfaces with minimal effort required.

All of the steps involved in the terminology process are performed with two very different types of systems in mind, and thus creating synergies in the process that are not found in conventional workflows.

The terminology that must always be generated in an MT context can be seamlessly implemented in numerous other contexts such as corporate terminology databases or web-based terminology solutions.

The return on investment for the installation and maintenance of an MT system is significant if the features of the system can be used for purposes other than those inherent to the system itself. And if the additional field of application of the system provides cost-related benefits and substantial efficiency and quality improvement, as would be the case with systematic terminology work, then the success of the implemented processes will far outweigh the investment within a very short period of time.

References

RBMT: Lucy LT by Lucy Software and Services GmbH

Text analytics system 1: acrolinx IQ by acrolinx GmbH

Text analytics system 2: tfTerm by text&form GmbH

Central Infrastructure: pentübridControl by text&form GmbH