On Integration of Terminological Data in Translation Systems

Signe Rirdance, Andrejs Vasiljevs

Tilde, Latvia

signe.rirdance@tilde.lv, andrejs@tilde.lv

Abstract

The current translation practice demonstrates lack of integration support between the traditional desktop translation tools and the rich terminological data available on the internet. This article sets the background for development of a new layer of web-based translation tools for automated translation of multilingual terminology, bridging the gap between translation tools and environments and internet term banks. It analyses the experience gained during the EuroTermBank project that proposes solutions to a number of challenges in integration of term banks with translation tools, such as the federation approach to interlinking term banks and the entry compounding approach for visual representation of multiple overlapping terminology entries. The article propones a standards-based approach to ensure data compatibility, and identifies the requirement to support terminology sharing on an interoperable level.

Index Terms: term bases, translation tools, terminology sharing

1. Introduction

In today's translation practice, a significant gap exists between the traditional desktop translation tools and the terminological data available on the internet. Translators spend from 30 up to 60% of total translation time on terminology research, therefore it is vital to ensure that they can use all the required terminology resources in the right format and in the right environment. Currently, translators spend a lot of time inefficiently, searching and processing information from multiple sources and changing its format to the one they require in their work environment. Spending time on technical aspects instead of focusing on true terminology research results in cost inefficiencies and reduced translation quality. Moreover, translation practice often involves redundant work in identifying, creating or compiling the same terminology over and over again, by various translators.

To reach a new level of translation productivity, a layer of new tools and technologies is required that 1) enables consolidation and integration of dispersed terminology resources; 2) provides online access to consolidated multilingual resources through internet term banks; 3) provides tools that connect specific translation environments with terminology resources on the internet; 4) introduces standards that enable terminology interoperability, sharing and reuse.

2. Discussion

This article sets the background for defining a toolset required for integration of terminological data into translation environments. It shortly reviews the state-of-the-art in commercial translation systems and analyzes the experience and best practices from the EuroTermBank project in consolidating diverse terminology resources. It wraps up by introducing a layer of tools being developed to support integration of multiple term banks with a diverse set of typical translation environments.

2.1. Terminology and translation systems

Computer tools and technologies that serve the purpose of assisting human translation are commonly known as CAT or Computer Assisted Translation software, also sometimes referred to as MAHT (Machine Assisted Human Translation) [1].

The basic environment assisting human translation is text processing applications that typically provide very basic CAT features like spell-checking and grammar checking; no terminology support is provided. Microsoft Word provides the additional function of searching in predefined reference resources and sites, provided in English and some other major languages.

Since end of 1980s, translation memory (TM) tools have been developed that utilize alignments or linkages between source and target texts. Often, translation memory tools include a terminology management module that enables the translator to search automatically in a given terminology database for terms appearing in a document, however, this function is limited to searching in a proprietary terminology base format. Examples of TM tools and their terminology management modules are: SDL Trados and SDL MultiTerm, Wordfast, DeJa Vu, Star Transit and others.

There are, however, some major drawbacks of these tools regarding handling of terminology for translation.

The most widely used translation environment tool's terminology module, MultiTerm is a full-fledged terminology management application, and as such, its complexity by far exceeds the complexity required by majority of translators. Translators using SDL Trados usually do not exploit the potential of MultiTerm and refrain from creating and using terminology.

Unlike in translation memory handling, providing efficient terminology recognition requires language-specific support to match inflected term forms with regular forms in the dictionary. Translation tools on the market provide morphology support for only few major languages, which is insufficient in the global multilingual environment.

Most translation tools used by freelance translators provide no support for internet resources or internet-based communication with the language workers' communities on the internet. While server-based work using embedded terminology workflows is supported in systems used by multinational corporations and organizations, their cost is prohibitive for the average industry practitioner.

2.2. Terminology consolidation in EuroTermBank

The goal of EuroTermBank project [1] is to facilitate terminology data accessibility and exchange, by collecting,

consolidating and disseminating dispersed terminology resources through an online terminology data bank. The initial focus of EuroTermBank was to contribute to improvement of the terminology infrastructure in the selected new European Union member countries (Latvia, Lithuania, Estonia, Poland, Hungaria) but project expands its activities to other countries in EU and beyond.

The objective of EuroTermBank is to become the leading site for integration of multilingual terminology resources into the central EuroTermBank database or interlink them via EuroTermBank as a central gateway and single point of service. The data bank works on a two-tier principle – as a central database and as an interlink node or gateway to other national and international terminology banks. Data exchange mechanisms have been developed to establish term import, export and exchange with other terminology databases. EuroTermBank multilingual terminology base is freely accessible online at http://www.eurotermbank.com.

2.2.1. Federation approach in consolidation of term banks

Federation is a new concept in linking portals and also data repositories, which goes far beyond the establishment of pointers or links, but reaches out to the level of semantic interoperability of data and data structures. Especially terminology and other kinds of structure content can be made to enable interoperability in the form of network(s) of federated databases [3].

Semantic interoperability and implementation of the federation principle are essential for the next level of integration of terminological data in translation environments. Consolidation of content, application of unified standards and semantic interoperability significantly eases the task of providing the layer of tools required to seamlessly integrate diverse term banks with diverse translation environments.

Today, however, terminology resources on the internet remain fragmented across diverse term banks and terminology projects. While it is clear that national or institutional terminology can be best identified in the terminology database of the respective institution, a number of user scenarios require consolidation on a multilingual and multinational scale. EuroTermBank not only stores all available terminology content in its database, but also acts as a gateway providing unified access to multiple remote terminology databases.

To ensure the viability of the federated system of terminology databases, inclusion of a termbank in this federated model requires it to be independently supported and maintained both institutionally and technically.

Within EuroTermBank, the mechanism that enables federation of external databases is called interlinking. Interlinking an external database to EuroTermBank enables users to query the external database from EuroTermBank web interface. It is implemented by connecting to the external resource through a web service, ensuring platform-independent interoperable machine-to-machine interaction over a network. Communication is done using XML messages that follow the SOAP-standard, a protocol for exchanging XML-based messages over computer networks, normally using HTTP.

Important steps towards a federated interoperable model of terminology management within an international organization are taking place in ISO. The ongoing project of developing the ISO Concept database envisions a federated approach to development and maintenance of content, as well as public access to ISO terminology, in the form of ISO electronic dictionary [4].

2.2.2. Entry compounding in consolidation of terminology content

Automated entry compounding is an innovative mechanism proposed by EuroTermBank in unification of potentially matching terminology entries from different resources. This novel concept is a cost-efficient solution to consolidated representation of terminology resources. In regards to translation practice, it carries important implications for new web-based approaches to efficient handling of terminology entries from multiple sources, which is a typical translator's scenario that has limited or no support in translation tools.

EuroTermBank data structure is modeled according to concept-oriented approach to terminology. Terminology entry denotes an abstract concept that has designations or terms as well as definitions in one or more languages. If terminology bank contains entries coming from different collections and designating the same concept, there is an obvious interest to merge them into one unified multilingual entry.

For example, if we have term pair *EN tree – LV koks* coming from a Latvian IT terminology resource and another term pair *EN tree – LT medis* from a Lithuanian IT terminology resource we may want to join these two into unified entry *EN tree – LV koks – LT medis*. Such multilingual entry allows to get correspondence between language terms that are not directly available in any terminology resource (in our example, the new term pair *LV koks – LT medis*).

However, merging entries just on the basis of a matching term in one language that is common for these entries will lead to many erroneous term correspondences, due to the frequent ambiguity of terms among subject fields or much rarer cases of ambiguity in the context within one subject field. The only error-free method for merging entries is evaluating whether these entries denote the same concept, however, it is often impossible or very expensive to make comparisons of cross-lingual terminology concepts, especially taken large databases like EuroTermBank thank contain over 1.5 million terms. Therefore, we propose a practical solution by introducing terminology entry compounding, which is an automated approach for matching terminology entries based on available data.

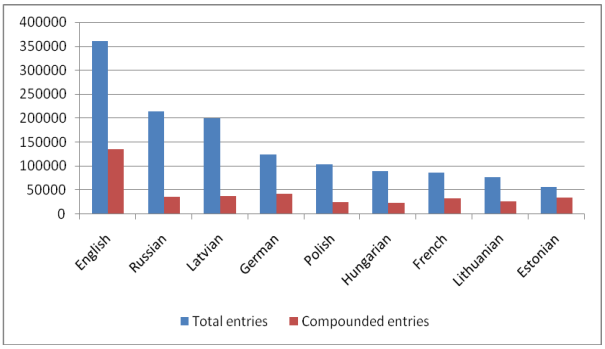

Figure 1. **Total and compounded entries per major languages of EuroTermBank.**

Entry compounding solves the problem of visual representation of multiple potentially overlapping term entries that are present in a consolidation of a huge number of multilingual terminology sources. At present, the

EuroTermBank database contains over 585,711 term entries with more than 1,500,500 terms. When applying entry compounding, over 135,000 or 23% of entries get compounded (see Figure 1). Hence entry compounding is a considerable aid for the user in finding the required term, for example, in the translation scenario between language pairs for which term equivalence is not established in existing collections.

Further work is planned on evaluating and improving entry compounding results within EuroTermBank, by applying corpus-based context analysis methods.

2.2.3. Application of standards

Integration of a multitude of term banks with a multitude of translation environments is only possible by rigorous implementation of applicable international standards. EuroTermBank project proposes a standards-based approach that is based on the best practice assessment and methodology recommendations developed by the EuroTermBank Consortium [5]. It includes a variety of aspects, from describing terminology collections to defining the data model and ensuring a unified data exchange format.

One of the major tasks in terminology data consolidation is identification and description of terminology resources. Due to a large number of resources to be described and different organizations in several countries involved it is important to use a common format for resource description. For this purpose we propose to use TeDIF format (Betz, Schmitz, 1999).

The Terminology Documentation Interchange Format TeDIF [6] was developed in the framework of the TDCnet project – European Terminology Documentation Centre Network, co-funded by the EU Commission. The TeDIF format was developed with the purpose to establish a common format for bibliographical and factual data related to terminology.

TeDIF provides means to describe bibliographical data like literature (serials, monographs, articles, journals, theses, etc.) and term collections (printed dictionaries, glossaries, thesauri, classifications, terminology databases, etc.).

EuroTermBank experience demonstrates applicability of TeDIF as a standard for terminology resource meta-data description and we recommend this format for other similar activities.

EuroTermBank is an early implementation [7] of the TBX (TermBase eXchange) standard, which enables a standard terminology exchange between term banks. The TBX standard, which is on its way to become an ISO standard in 2009, is based on three ISO standards: ISO 12620, ISO 12200 and ISO 16642. ISO 12620 defines data categories to be used for terminological data storage either in digital or printed format, while ISO 12200 defines MARTIF, an SGML interchange standard for interchange of terminological data and ISO 16642 that defines TMF, the terminological markup framework.

2.2.4. Terminology sharing

A number of emerging areas of activity will require the new tools layer that, first and foremost, connects term banks and translation environments, but also allows publishing and distributing terminology resources that comply with the standards required for this connectivity. Thus, terminology sharing is a phenomenon that involves sharing of non-confidential, non-competing and non-differentiating terminology across various actors – individuals along with companies and language service providers, with the goal to consolidate and promote accessibility to multilingual terminology per vertical industries [8]. Terminology sharing involves returns from streamlined industry terminology, by ensuring reuse of existing terminology assets. For those who share their terminology, it is a way of promoting and disseminating one's well-established terminology, possibly even to the level of de facto industry standard terminology. However, to reap full benefits from the shared terminology, it is essential to ensure integrated access to these terminology resources in translation environments.

3. Conclusions

Currently, the richness of terminology resources on the internet does not translate into the expected increased productivity and quality levels of translation work, and the area of providing tools for integration of terminological data in translation systems is relatively new. Therefore, it is important to establish a few basic principles that enable easier integration of term banks with the variety of translation environments.

The federation principle interlinks independently maintained term banks and provides a consolidated access point for terminology searches by human or machine users.

Terminology entry compounding provides a method that, from the translator's perspective, cuts translation inefficiencies in looking up a multitude of resources.

An underlying principle is a standards-based approach in developing term banks and any internet terminology resources, to ensure that the data can be easily exchanged in various terminology exchange and terminology sharing scenarios.

A new layer of tools and technologies that integrate translation environments with terminological resources on the internet is required to significantly enhance the current productivity of human translation.

4. Acknowledgements

Many thanks to colleagues in all EuroTermBank project partner organizations: Tilde (Latvia), Institute for Information Management at Cologne University of Applied Science (Germany), Centre for Language Technology at University of Copenhagen (Denmark), Institute of Lithuanian Language (Lithuania), Terminology Commission of Latvian Academy of Science (Latvia), MorphoLogic (Hungary), University of Tartu (Estonia), Information Processing Centre (Poland). EuroTermBank Consortium would also like to acknowledge and thank the European Union eContent program for supporting the EuroTermBank project as well as support from EU Social Fund.

5. References

[1] Hutchins J., "Current commercial machine translation systems and computer-based translation tools: system types and their uses", International Journal of Translation**,** vol.17, no.1-2, pp. 5-38, Jan-Dec 2005.

[2] Vasiljevs A., Skadins R., "EuroTermBank terminology database and cooperation network", proceedings of the Second Baltic Conference on Human Language Technologies, Tallinn, pp. 347-352, 2005.

[3] Galinski C., "New ideas on how to support terminology standardization projects", eDITion, 1/2007.

[4] Weissinger R., ”Integrating Standards in Practice”, *ISO Concept Database* presentation, 10th Open Forum on Metadata Registries, New York, July 2007.

[5] Auksoriute A. et al, “Towards Consolidation of European Terminology Resources. Experience and Recommendations from EuroTermBank Project”, Riga, 2006.

[6] Betz A., Schmitz K.-D., “The Terminology Documentation Interchange Format TeDIF”, Terminology and Knowledge Engineering TKE'99, Innsbruck, Wien, pp. 782-792, 1999.

[7] Vasiljevs A., Liedskalnins A., Rirdance S., “From Paper to TBX: Processing Diverse Data Formats for Multilingual Term Bank”, proceedings of the Third Baltic Conference on Human Language Technologies, Tallinn, 2008 (will be published).

[8] Rirdance S., “IP vs. Customer Satisfaction: EuroTermBank and the Business Case for Terminology Sharing”, The Globalization Insider, LISA, 6/2007.