Terminology Extraction as a Tool for MT Output Assessment and Improvement

Jean-François Richard

President, Terminotix

Montréal, Qc, Canada

jfrichard@terminotix.com

Abstract

The present paper proposes the use of a Terminology Recall Index (TRI) calculated on retaining nominal groups' frequencies and stemming info only. Though this paper proposes to demonstrate the utility of a TRI calculation between a human translated document and neural machine translated document, it also attempts to demonstrate that a broader use of the TRI calculation has many other surprising applications inside a linguistic service's translation workflow.

1 Introduction

Most often, metrics currently used for evaluating the percentage of similarity between two documents in the same language are based on various types of edit distance algorithms. The edit distance calculation represents the effort required to make two documents identical. The bilingual evaluation understudy (BLEU) score is also based on an edit distance algorithm. This paper will first demonstrate, from a linguistic point of view and as explained in Section 2, why this metric is problematic.

It will then propose a more accurate metric with which linguists[35] can evaluate the edit distance between two target versions of a given document using semantic term extraction algorithms: the terminology recall index (TRI). It will also propose that an automated TRI report be generated automatically as part of the translation workflow process.

This paper will cover in detail the TRI applications identified to date that could be integrated into a translation workflow with the primary goal of automating the terminology quality assurance (QA) process while improving productivity.

In conclusion, the paper will describe the currently identified limitations of this approach and explore possible avenues for research and development in order to remedy these limitations.

2 Context and assumptions

During the translation phase[36], searching tasks take between 30 and 40 percent of linguists' time[37], as explained under Subsection **Erreur ! Source du renvoi introuvable.**. The terms that linguists search for are mainly noun groups, and the terminology databases maintained internally by language services are also mainly composed of noun groups, as explained under Subsection 3.2.

[35] Linguists are employees in language services departments. Their roles include translator, post-editor, editor, reviser, coordinator, proof-reader, and terminologist.

[36] A translation request can go through different processes, such as pretranslation, post-editing, editing, proofreading, bilingual revision, typesetting, desktop publishing, bitext creation, etc. Requesters submit their requests either via a translation workflow platform or by email.

[37] Depending on the nature and complexity of the content, searching time can be as much as 50% and as little as almost nothing.

3 More than one day per week spent searching

For quality and consistency purposes, linguists often search for terms in many different sources. Depending on his or her role, a linguist may use the following sources:

- Glossaries and dictionaries, for finding definitions, synonyms, etymology, etc.
- Internal and/or online terminology databases, for finding specific terms for a specific domain or client
- Bitext search engines, for finding terms and their translations in context
- Translation memories, for finding bilingual concordances of a term
- Full-text search engines, to see how a term is used in the original language
- Full-text search engines, for finding previously translated documents or reference materials

In all these cases (excepting that of the final bullet point[38]), the part of speech most often searched for is the noun group. It is also the most time-consuming part of speech to process. Although other parts of speech are also searched for, they do not take up as much time as noun groups. With the exception of idiomatic expressions[39], other parts of speech are not challenging for linguists.

In some language services[40] departments, if a search garners a satisfactory result, linguists will update the internal terminology database (TB), whether upon instruction or of their own initiative.

3.1 Noun groups in terminology databases

Most internal language services departments in organisations maintain an internal TB. This database mainly comprises records for noun groups. The same linguists who search within this TB also create and maintain term records, as described above under Subsection **Erreur ! Source du renvoi introuvable.**.

3.2 A more accurate metric

Based on these observations, as covered under Subsections **Erreur ! Source du renvoi introuvable.** and 3.1 above, and as the BLEU score treats all parts of speech equally, whether low-weight semantic entities (articles, conjunctions, prepositions, etc.) or high-weight semantic expressions (noun groups, etc.), it is safe to assume that a more accurate metric based on noun groups only would be useful to linguists.

4 Term extraction for noun groups

This section describes how a semantic term extraction engine, required for calculating TRI, has been developed by combining existing term extraction algorithms with existing part-of-speech-tagger algorithms.

[38] The reference searching task is usually done by coordinators and does not require any linguistic skills.

[39] Expressions such as “he’s out to lunch,” “the proof is in the pudding,” and “I don’t have a dog in that fight,” which are challenging for machine translation engines.

[40] Not all language services departments maintain or even use a centralized TB. This is especially true for small departments. Often, linguists search for the same terms over and over, increasing overall time spent searching. Sometimes, terminology is maintained in a Word or Excel document.

4.1 Existing term extraction engines

Most existing term extraction engines are based on a statistical model, using lists of "stop words" for noise reduction. Term extraction engines are somewhat useful for identifying patterns of co-occurring words[41], sorted by frequency. However, besides some basic stop word noise reduction algorithms, statistical term extraction remains "noisy" and cannot generate high quality noun group extractions. Some term extraction tools[42] provide user interfaces to manually remove noise entries.

4.2 Existing tagger engines

Existing taggers[43] are able, among other things, to a) Identify parts of speech and modifiers for each word in a given sentence; b) to build a sentence grammar tree[44]. Since taggers are more language sensitive, supported languages are more limited than for term extractors. Though grammatical information is usually well extracted, it does not provide clearly extracted, useful to linguists, "noun groups".

4.3 Integrating taggers into term extractors

By "plugging" the multiple word phrase outputs, generated by a regular statistical term extraction, into a tagger engine, a "semantic term extraction" engine can be developed and integrated. This first and fast statistical pass helps in reducing basic stop word noise and to retain multiple words only. During the second pass, a tagger is used for identifying part of speech patterns enabling to mark an extracted phrase as a potential noun group.

5 Proposed TRI formula

This section presents a proposed formula for calculating TRI against many sources like MT output, internal database, source document, human translation. First, we calculate a Terminology Frequency Sum Index (TFSI) by summing each frequency for each noun groups extracted in a given document d:

$$\mathrm{TFI}(d) \ = \sum_{k=1}^{n} F(k)$$

Where F is the frequency of an item at position k in a list of n extracted noun groups. The following table gives an example of a TFI calculation:

[41] For basic disambiguation purposes, only multiple word phrases are retained. Lemmatized single words are highly subject to polysemy.

[42] SDL MultiTerm Extract, SynchroTerm, CrossMining, MultiTrans Term Extractor, etc.

[43] Spacy, TreeTagger.

[44] Generative grammar theory by Noam Chomsky.

Noun groups	**Frequencies**
états financiers	15
anomalie significative	11
associé responsable	6
règle de déontologie	5
société de services	3
modalité d’application	3
opinion avec réserve	2
seuil de signification	2
cabinet membre	2
TFSI(d)	49

Table 6: Example of a TFSI calculation from a human translated document

In order to calculate a TRI between two documents of the same language, we calculate a TFSI for the first document (d) and a TFSI for the second document (d^1). Suppose we want to get the TRI against the TFSI(d) obtained in Table 1:

Noun groups	**Frequencies**
états financiers	14
anomalie significative	10
partenaire responsable	6
règle de déontologie	5
entreprise de services	3
modalité d’application	3
opinion avec réserve	2
seuil de signification	2
bureau membre	2
TFSI(d^1)	47

Table 2: Example of a TFSI calculation from a machine translation document

We then build a list of noun groups present in d crosschecked with d^1 by using the following TRI global formula:

$$\mathrm{TRI}(d, d') = \forall F(d), F(d^1) \in (\mathrm{TFI}(d) \cap \mathrm{TFI}(d^1)) \colon (\sum_{k=1}^{n} \mathrm{Min}(F(k), F'(k))\ ) / (\mathrm{Max}(\mathrm{TFSI}(d), \mathrm{TSFI}(d')) * 100$$

By using the data from Table 1 and Table 2, we calculate that the Terminology Recall Index between d and d’ is 73%:

TFI(d) ∩ TFI(d^1)	**F(*k*)**	**F’(*k*)**	**Min(F(k),F’(k)**
états financiers	15	14	14
anomalie significative	11	10	10
règle de déontologie	5	5	5
modalité d’application	3	3	3
opinion avec réserve	2	2	2
seuil de signification	2	2	2
TFSI(d, d’)	49	47	36
TRI(d, d’)			**73%** (36 / 49 * 100)

Table 3: Example of a TRI calculation between d and d’

6 TRI applications

This section lists possible TRI integrations and their respective applications.

6.1 TRI integration with MT

Here are some examples of integrating TRI with MT engines: 1) Calculate the TRI between a human translation and a machine translation of the same document, allowing linguists to obtain a general TRI and identify which noun groups were and were not properly translated by the MT engine. 2) Use the TRI for analysis purposes when running bench tests, or when comparing MT engines. 2) Employ the TRI to identify which MT engine should be used for a given source document by sending only the most frequent noun groups to the MT engines and cross-checking the output against an internal terminology database, or against the terms extracted from the human translation.

6.2 TRI integration with source documents

Here are some examples of TRI applications for documents prior to translation: 1) Calculate the TRI by cross-checking against existing terminology in order to identify which terms are and are not present in the internal terminology database. The list of known terms present both in the document and the term database is called a “job glossary” extraction, while the list of unknown terms in the same document is called a “unilingual term candidate” extraction.

6.3 TRI integration with target documents

Generating a TRI that compares the final target document with internal terminology, or with a job glossary initially generated from the source document, can provide linguists with the following: 1) A list of terms present but not properly translated; 2) A list of terms not present at all; 3) A TRI for automatic workflow notifications, or for automatically feeding a database of term candidates. Linguists in charge of terminology could take a few moments to review the weekly TRI reports, which would become a tool for monitoring the quality of terminology.

7 Limitations

The two limitations for deploying TRI are as follows: 1) Since the TRI retains noun groups of two or more words only, for co-occurrence semantic reasons, it is less useful in subject matter fields in which single-word noun groups are more frequent; 2) The list of available taggers for languages other than English, French and German are more difficult to find and/or integrate into code.

8 Future developments in terminology automation

Apart from the two limitations mentioned in Section 7 above, for which remedies should be achievable in the very near future, bilingual noun group extraction and/or translation spotting could be integrated by automatically identifying suggestions for the unilingual term candidates TRI list, which would be generated upon the initial receipt of a document for translation. This list of bilingual term candidates could also be automatically added to a terminology database with the same name. Then, a quick review and validation/rejection task could be performed on a weekly basis.

References

Lynne Bowker and Jairo Buitrago Ciro. 2012. Machine Translation and Global Research: Towards Improved Machine Translation and Global Research.

Joachim Wermter. 2008. Collocation and Term Extraction Using Linguistically Enhanced Statistical Methods.

Douglas Robinson. Becoming a Translator: An Introduction to the Theory and Practice of Translation. Page 35.

Youakim Badr, Richard Chbeir, Ajith Abraham. 2010. Emergent Web Intelligence: Advanced Semantic Technologies. Page 251.

Dan Melamed. 2001. Empirical Methods for Exploiting Parallel Texts.