Defining a Gold Standard for the Evaluation of Term Extractors

Gabriel Bernier-Colborne
Observatoire de linguistique Sens-Texte
Université de Montréal
E-mail: gabriel.bernier-colborne@umontreal.ca

Abstract

We describe a methodology for constructing a gold standard for the automatic evaluation of term extractors, an important step toward establishing a much-needed evaluation protocol for term extraction systems. The gold standard proposed is a fully annotated corpus, constructed in accordance with a specific terminological setting (i.e. the compilation of a specialized dictionary of automotive mechanics), and accounting for the wide variety of realizations of terms in context. A list of all the terminological units in the corpus is extracted, and may be compared to the output of a term extractor, using a set of metrics to assess its performance. Subsets of terminological units may also be extracted, due to the use of XML for annotation purposes, providing a level of customization. Particular attention is paid to the criteria used to select terminological units in the corpus, and the protocol established to account for terminological variation within the corpus.

Keywords: term extraction, evaluation, annotated corpora, gold standard

1. A Gold Standard for Term Extraction

Terminological resources are compiled using various technologies, but few of these technologies have been evaluated from the point of view of their contribution in a specific terminological setting. Since methodologies for compiling these resources are increasingly corpus-based, one of the main tools is the term extractor. Term extractors are used for compiling specialized dictionaries (L'Homme, 2008), ontologies (Biébow & Szulman, 1999) and back-of-the-book indexes (Nazarenko & Aït El Mekki, 2005). This paper describes a proposal for the definition of a gold standard for automatically evaluating term extractors, an important step toward establishing a much-needed evaluation protocol for term extraction systems.

Term extractors are tools designed to retrieve specialized terms from running text, which play a role in a variety of applications, such as terminology, thesaurus building, document indexing, technological watch and ontology development. Like all language technologies, the design and improvement of term extractors requires that developers evaluate these systems.

When new extraction techniques are introduced, attempts are usually made to measure their performance, but how this evaluation is undertaken varies greatly and is often not described in much detail. In some cases, extractors are evaluated by manually scanning their output. In others, extractor output is compared to some sort of term list, but this reference is seldom given much attention in the literature.

This lack of a standardized evaluation protocol has motivated some researchers in the field (L'Homme et al., 1996; Sauron, 2002; Vivaldi & Rodríguez, 2007; Nazarenko & Zargayouna, 2009) to make proposals for such a protocol. Large-scale evaluation efforts have been undertaken in the form of campaigns or workshops, such as ARC A3 and CESART (Timimi & Mustafa el Hadi, 2008) as well as NTCIR-TMREC (Kageura et al., 2000), but these efforts pale in comparison with those made in other branches of NLP (Nazarenko et al., 2009).

If a standardized automatic evaluation protocol is to be established, a gold standard must be defined. To this end, we propose a fully annotated corpus, accounting for the wide variety of realizations of terms in context. This standard is the reflection of a specific terminological setting, namely compiling a specialized dictionary; other applications would necessarily derive a different set of terms.

This gold standard can also be customized, due to the use of XML for annotation purposes. Combined with a set of metrics, such a standard will enable an automatic evaluation of the performance of term extractors, which would be helpful in assessing the performance of a particular system given a specific setting, or comparing different techniques. It would also allow developers to fine-tune their systems by measuring how a given component affects the overall output or how a change in design affects performance (Popescu-Belis, 2007: 77).

2. Specific Problems in Term Extraction

Term extraction raises challenges that are not found in other NLP technologies:

- The notion of "term" is linked to a specific application, as users of a term extractor have different needs in accordance with their professional activity (Estopà Bagot, 2001), be it knowledge organization, indexing or specialized lexicography. Thus, the relevance of terms depends on the task at hand.

- The use of terms in context involves various phenomena that modify their normal structure and can make the identification of term boundaries

difficult. These include coordination of complex terms (e.g. *room temperature vulcanizing and anaerobic sealants*), embedding of terms or other elements within compound terms (e.g. *inline (placed in the fuel line) filter*) and anaphoric references (e.g. using *tank* when *fuel tank* has been used previously).

- Concepts may be denoted by more than one term, and terms are subject to various kinds of variation, such as regional variations (e.g. *gearbox*/*transmission*), spelling variations (e.g. *disc brake*/*disk brake*), syntactic variations (e.g. *piston head*/*head of the piston*) and acronyms, which adds a level of complexity to term selection. These variants must be encoded in terminological resources to allow language technologies to recognize them.

Each of these factors also has an impact on the evaluation of term extractors. First, since the ideal set of terms provided by an extractor varies according to the task involved, the reference used to evaluate an extractor must be compiled in accordance with a specific application. With this in mind, we chose the compilation of a specialized dictionary as the application guiding the selection of terms; thus, the gold standard is meant to reflect the work of a terminologist.

The next section will focus on the factors that make term selection and term boundary identification difficult, and how they were dealt with during annotation of the corpus.

3. Annotating the Corpus

The corpus that was annotated and is used to establish the gold standard set of terms -- which will also be used as the test corpus for evaluating term extractors -- consists of three manuals on automotive mechanics, containing some 224,159 tokens. A set of guidelines for selecting terms was established, which includes some of the term selection criteria described by L'Homme (2004: 64--66).

First, units must convey a meaning that is related to the chosen subject field. Units that are morphologically related to previously selected terms (e.g. *cooling pump* and *coolant pump*) are also valid, as long as they are also semantically related. Units that share a paradigmatic bond (e.g. synonymy, meronymy, etc.) with valid terms are also likely candidates. The criteria set out by L'Homme concerning predicative units were not used, as it was decided that only nouns and noun phrases were eligible, since most of the concepts that should be included in a dictionary of this field are denoted by nouns.

Moreover, only units of maximum length are selected, such that terms embedded within terms are not tagged as such.

Regional variations, spelling variations and acronyms are included, and the type of variation is specified in the term bank (see Section 4).

More general, or thematic, guidelines were also followed, based on the idea that the application guiding the selection of terms was the production of a specialized dictionary that focuses mainly on the structure of an automobile. Accordingly, terms denoting parts, types of cars and products that a car needs to work are included, whereas terms denoting damages or units of measurement are excluded.

All terms considered relevant according to these guidelines are tagged within the corpus in XML format. These tags serve not only to segment the text into terms and non-terms, but to identify them using a unique identification number and describe certain features of the terms (simple or compound, types of variation), as shown if Figure 1. The selection and tagging process was entirely manual -- term extractors were not used to pre-process the text.

```
One <term id="1" type="s">transaxle</term>
design is the <term id="2" type="c">continuously
variable transmission</term> (<term id="3"
type="a">CVT</term>).
```

Figure 1: Tagged text (with simplified tags).

Rules were established for cases where term segmentation is not so straightforward, as compound terms may be truncated for various reasons.

Coordinated compound terms (e.g. *intake and exhaust valves*) are tagged separately. In compound terms that are disjoined by punctuation marks, embedded terms or paraphrases (e.g. all wheel drive (AWD) systems), any linear sequence that corresponds to a term or part of term is tagged as such, and extraneous elements are excluded whenever possible, as shown in Figure 2. Anaphora can also result in compound term truncation, and these forms are tagged as well.

In all cases, shortened forms are linked to the base term, as described below, and their tags contain an attribute indicating that they are variations of some base term.

```
Many manufacturers have introduced <term
str="coord">full time FWD drive</term> or
<term str="coord">all wheel drive</term>
(<term str="disj">AWD) systems</term>
```

Figure 2: Disjoined compound terms (simplified tags).

4. Building the Term Base

The tagged terms are then entered, in their lemmatized form, in a separate term base, in which each term has a record, as shown in Figure 3.

Each sense of a polysemous term receives its own record and identification number. This number not only establishes a distinction between senses, but also allows for easy retrieval of term occurrences from the corpus. Records include a definition, generally adapted from a term base or specialized dictionary, which allows the annotator, as well as any future users of the term base, to obtain the term's meaning, and distinguish between polysemous terms.

Also included in the record is information concerning synonymy and term variation. Synonyms and variations are linked together, all forms pointing to one base term, which is chosen by looking for the headword most often used in dictionaries and term bases, and for the term that has the highest frequency in the corpus. Compound terms that are truncated for the reasons described above (coordination, embedding, anaphora) are given their own record, including a link to the base term. If the base term did not occur in the corpus, the term is "reconstructed" and given a record in the term base.

ID	307
Lemma	EGR valve
Variation type	Acronym
Base term	exhaust gas recirculation valve
Definition	A valve that regulates the flow of exhaust gas into the intake manifold.

Figure 3: Part of the record for the term *EGR valve*.

Inspection of the term base reveals some interesting properties. The term base contains 5489 records, more than half of which, interestingly, are not base terms: 1257 are synonyms of a base term, 1447 are truncated forms of compound terms, and 55 are acronyms. 174 terms were reconstructed from truncated compound terms. Furthermore, of the 23 terms that have a frequency greater than 100, none is compound, if we exclude the base terms that two variations derive from. The corpus contains 28,658 term occurrences, yielding an average term frequency of 5.22, and contains 2,656 hapax legomena -- regarding these figures, it is important to remember that the different meanings of polysemous terms are considered separately. Although this is outside the scope of this paper, a clearer picture of the distribution of terms in the corpus might be provided if frequencies were calculated not only on individual terms (or senses), but also on sets comprising a term and its variations.

The term base can be used as is and compared to the output of a term extractor using a set of metrics. These could be traditional metrics, such as precision and recall, or other metrics that have been proposed for term extraction evaluation (Nazarenko et al., 2009). It is also possible to extract subsets of the term list, for example by excluding uniterms or specific types of terminological variations. This can be easily accomplished using XSLT, and produces a customized term list for the purposes of evaluation.

5. Conclusion

In this paper, we have described a methodology for constructing a gold standard for the automatic evaluation of term extractors. Particular attention has been paid to term selection criteria and term segmentation, as well as the processing of terminological variations.

The gold standard was built by annotating a corpus on automotive mechanics in accordance with a specific application, namely compiling a specialized dictionary. Extensions of this work might include annotating terms in a corpus in accordance with more than one application (ontology development, document indexing, etc.), which would allow evaluators to measure the relevance of extractor output to different applications.

Using the gold standard to evaluate a term extractor is fairly straightforward. The tags are removed from the corpus, which then serves as the test corpus: it is fed to a term extractor, and the output is compared to the standard using an appropriate set of metrics. This enables the performance of a term extractor to be assessed automatically, an important step toward establishing a standardized automatic evaluation protocol for term extractors.

6. Acknowledgements

The work presented in this paper is supported by the Quebec funding agency *Fonds de recherche Société et culture* and the Social Sciences and Humanities Research Council of Canada.

The author also wishes to thank Patrick Drouin and Marie-Claude L'Homme for their helpful suggestions regarding this paper.

7. References

Biébow, B., Szulman, S. (1999). Terminae: A Linguistics-Based Tool for the Building of a Domain Ontology. In *Proceedings of the 11th European Workshop, Knowledge Acquisition, Modelling and Management (EKAW 99), LNAI 1621*. Berlin: Springer Verlag, pp. 49--66.

Estopà Bagot, R. (2001). Les unités de signification spécialisées : élargissant l'objet du travail en terminologie. *Terminology*, 7(2), pp. 217--237.

Kageura, K. et al. (2000). Recent advances in automatic

term recognition: Experiences from the NTCIR workshop on information retrieval and term recognition. *Terminology,* 6(2), pp. 151--173.

L'Homme, M.-C., Benali, L., Bertrand, C., Lauduique, P. (1996). Definition of an Evaluation Grid for Term-Extraction Software. *Terminology*, 3(2), pp. 291--312.

L'Homme, M.-C. (2004). *La terminologie: principes et techniques*. Montréal: Presses de l'Université de Montréal.

L'Homme, M.-C. (2008). Le DiCoInfo. Méthodologie pour une nouvelle génération de dictionnaires spécialisés. *Traduire*, 217, pp. 78--103.

Nazarenko, A., Aït El Mekki, T. (2005). Building back of the book indexes. *Terminology*, 11(1), pp. 199--224.

Nazarenko, A., Zargayouna, H. (2009). Evaluating Term Extraction. In *Proceedings of RANLP 2009*, pp. 299--304.

Nazarenko, A., Zargayouna, H., Hamon, O., Van Puymbrouck, J. (2009). Évaluation des outils terminologiques : enjeux, difficultés et propositions. *TAL*, 50(1), pp. 257--281.

Popescu-Belis, A. (2007). Le rôle des métriques d'évaluation dans le processus de recherche en TAL. *TAL*, 48(1), pp. 67--91.

Sauron, V. (2002). Tearing out the Terms: Evaluating Terms Extractors. In *Proceedings of the Aslib conference Translating and the Computer 24*.

Timimi, I. Mustafa el Hadi, W. (2008). CESART : une campagne d'évaluation de systèmes d'acquisition de ressources terminologiques. In S. Chaudiron & K. Choukri (Eds.), *L'évaluation des technologies en traitement de la langue : les campagnes Technolangue*. Paris: Hermes science, pp. 71--91.

Vivaldi, J., Rodríguez, H. (2007). Evaluation of Terms and Term Extraction Systems: A Practical Approach. *Terminology*, 13(2), pp. 225--248.