TermoStat Exercise: Level I

Term extractors are used to identify important terms in a text or group of texts. Translators and other language professionals may find the results useful for getting an idea of the topic and content of a text, and as a starting point for terminological research. However, the process of automatically identifying key terms is based entirely on formal cues, and so the results can only be considered as *candidate* terms. Human evaluation of these results is always necessary to determine whether the units identified are really terms, and whether they are pertinent in a given context.

The identification of these candidate terms is generally based directly or indirectly on the frequency with which these units occur in a text or text collection, on their form, or on a combination of these two criteria. These approaches are generally referred to as *statistical*, *linguistic* and *hybrid*, respectively. Essentially, the assumption is that units that are particularly frequent, that correspond to standard patterns of term formation (e.g. NOUN + ADJECTIVE, NOUN + NOUN), or both, are likely to be terms.

TermoStat, a tool developed by Patrick Drouin at the Observatoire de linguistique Sens-Texte (OLST) at the Université de Montréal, is an online term extractor that uses a hybrid (i.e. statistical plus linguistic) method to identify candidate terms. It takes into account not only the structure of potential term candidates (using a program called a *part-of-speech tagger* to identify nouns and adjectives and complex structures that contain these items), but also the relative frequencies of these potential candidates in the text being processed (called the *analysis corpus*) and a very large collection of newspaper articles (called the *reference corpus*). This method allows TermoStat to find not only multi-word but also single-word candidate terms in a single extraction process.

Drouin refers to the calculation of the difference between the relative frequencies of the candidates in the analysis and reference corpora — the default score used to identify candidate terms — as *specificity*. TermoStat does nevertheless allow for the use of other measures to identify terms. Thus, this tool also makes it possible to compare the results of different approaches to term extraction in a given situation.

To obtain an accurate count of the number of occurrences of each candidate term, TermoStat uses a process known as *lemmatization*: it transforms inflected forms of words that appear in the corpora into base forms (e.g. transforming plurals of nouns such as *rights* to the singular *right*, and inflected forms of French adjectives to their masculine singular form, e.g. *ancestrale*, *ancestraux* and *ancestrales* to *ancestral*). Once this is done, all of the forms can be counted as occurrences of a single term instead of as separate terms. Because of this, TermoStat's results include two separate fields: the *candidat lemmatisé* column displays the form of a candidate term with its components restored to their base form, and the *variantes orthographiques* column displays the form actually found in the text itself.

You can learn more about these term-extraction techniques in *Initiation à la traductique* (L'Homme 1999), *Computer-Aided Translation Technology* (Bowker 2002), *La terminologie : Principes et techniques* (L'Homme 2004) and on the Web page of the

OLST (www.olst.umontreal.ca > *Ressources* > *Ressources informatiques pour la terminologie* > *Extraction de termes* > in the presentation *Extraction de termes : techniques courantes*). You can also learn more about TermoStat specifically in Drouin, P. (2003) “Term Extraction using non-technical corpora as a point of leverage”, published in *Terminology* 9(1): 99–115, and in a number of other articles on TermoStat, identified on Drouin’s Web page at www.mapageweb.umontreal.ca/drouinp/, as well as in Drouin’s doctoral thesis, also available on this page.

The exercises below will help you learn to:

- Use TermoStat; and
- Evaluate the results of the term extraction in terms of
  - the presentation of the results,
  - the kinds of units identified,
  - the form of the units identified,
  - the precision and recall of the extraction (i.e. the proportions of *noise* and *silences* in the results), and
  - the measures taken to improve the quality of the results, as well as their effects.

Getting ready

1. Create a sub-directory of the U: drive (also called **MY DOCUMENTS**). (For instructions, see the file *Creating a sub-directory* on the CERTT site (**ACCESS BY TOOL NAME > WINDOWSXP > CREATING A SUB-DIRECTORY**).)
2. Download an excerpt from the Wikipedia entry on in vitro fertilization from the CERTT site (**FILES TO DOWNLOAD > WIKIPEDIA IVF ENGLISH (EXCERPT)**).
3. Open the file and read through it to get an idea of its content. Identify some terms (between 5 and 10) that you consider to be pertinent in the text and make a note of them. Close the file.
4. Connect to the Internet (if necessary) and open the Web browser of your choice, for example Internet Explorer or Mozilla Firefox (using the shortcut on the Desktop or from the **START** menu).
5. Open the TermoStat Web page: http://olst.ling.umontreal.ca/~drouinp/termostat_web.

Extraction

1. Click on the **BROWSE** button and select your file.
2. From the *Language* dropdown list, select the language of your file.
3. From the *Test* dropdown list, select the test you would like to use to identify candidate terms. (Try using the default option, *Specificity*, first.)

4. Click on the **ANALYZE** button to start the extraction.

5. Once the analysis is completed, click on the **SEE RESULTS** button to display the candidate terms identified.

6. Look at the results to get an idea of the information about each candidate that TermoStat provides, and the presentation of the results:

   a. The left-hand column (*Fréquence*) indicates the frequency of the candidate-term in the analyzed file;

   b. The column to its right (*Candidat lemmatisé*) displays the lemmatized (i.e. base) forms of the candidate terms identified by TermoStat (that is, the word or series of words that form the candidate, after lemmatization);

   c. The next column (*Variantes orthographiques*) displays the form(s) of the candidates that were identified in the text itself (i.e. non-lemmatized forms as well as some spelling variants);

   d. The right-hand column (*Poids*) displays the candidates' specificity scores; by default the results are sorted in descending order according to this score. The specificity score is the product of the comparison of the candidate's frequency in your analyzed text and in the reference corpus. A high specificity score indicates that the candidate is specific to your text, i.e. that it is particularly frequent, and thus that it is considered to be a likely candidate term.

7. Click on a lemmatized candidate term to see the context(s) in which it was identified.

8. Click on the headers of the columns to sort the results by the different criteria. Observe the differences in the rank of the term candidates.

   a. Are some terms identified as good candidates according to one possible criterion (e.g. specificity) but not another (e.g. frequency)?

   b. What is the value of sorting the candidates according to the criteria *Candidat lemmatisé*? When or for what kinds of tasks might it be most useful?

9. Evaluate the results of the extraction according to the criteria outlined in the Introduction:

   a. What forms do the candidate terms take? To what part-of-speech classes do they belong? What kinds of structures do they have?

   b. What is the frequency of the candidate terms? Are they always fairly frequent?

   c. What is the precision of the results (i.e. how many of the candidate terms identified are really terms, in your opinion)? Is there any value for a translator in retaining candidates that do not strictly qualify as terms?

   d. What is the recall of the results (i.e. are all of the potentially interesting terms you identified when you read the text present in the results)? If they did appear on TermoStat's list, how high was their ranking and how does this compare with your opinion of the importance of the term? If they did not appear on TermoStat's list, why do you think they might have been left off?

   e. Can you identify any recurrent difficulties in the forms of the complex candidate terms identified by TermoStat? What is the source of these difficulties, do you think? Are the results useful despite these problems?

f. What effect does the process of lemmatization have on the presentation of the results? On the forms of the candidate terms? On the measurement of their frequencies? Did you identify any problems with this presentation?

g. Did you notice any other problems with the results of the extraction? Can you identify the sources of these problems by looking at the candidates and/or the contexts in which they appeared?

Wrapping up

1. To keep a copy of your results, copy the table created by TermoStat (e.g. by selecting it with your cursor and pressing **CTRL + C**) and paste it into a Word document or Excel spreadsheet (**CTRL + V**). Then save this file to the sub-directory you created at the beginning of these exercises. (You can also copy the file to a disk or USB key, or send it as an attachment to your e-mail.)

NOTE: If you click on the VIEW PREVIOUSLY SUBMITTED CORPORA button in the TermoStat interface, you can view a list of previously processed files. In future, you can thus consult the results of your extraction without having to re-process your file, just by finding the name of your file in the list. However, these files may disappear without notice, so if you think you will want a copy of your results, it is better to make one yourself.

You can also consult the results of other extractions. However, in order to interpret the results properly, it is of course necessary to know something about the original text. Note that you do not have access to this text through the TermoStat interface.

Questions for reflection

1. In light of your results, can you identify some strong and weak points of this extraction approach in various contexts?
2. Do you find TermoStat harder or easier to use, more or less user-friendly than some other extractors that you have tried? Why?
3. As explained in the introductory section, the *Variantes orthographiques* presented by TermoStat are principally the singular and plural forms of nouns (e.g. *right, rights*) or, in French, the inflected forms of adjectives (e.g. *ancestral, ancestrale, ancestraux, ancestrales*). What other forms of orthographic variation could exist in a text, and how would this affect the results of a term extraction system such as TermoStat?
4. Examine the list of candidates (and their contexts) carefully. Can you identify other challenges in NLP (e.g. among those described in L'Homme 1999, chapter 2) that may have implications for which candidates are selected? Which ones, and how may they affect the tool's performance?

Tutorial updated by Cheryl McBride