KWIC corpora as a source of specialized definitional information:
an pilot study

Antonio San Martín
LexiCon Research Group
Department of Translation and Interpreting
University of Granada
Calle Buensuceso 11, 18002 Granada (Spain)
http://lexicon.ugr.es/sanmartin
asanmartin@ugr.es

Résumé. Nous présentons une étude pilote visant à évaluer dans quelle mesure une liste de termes générée par l’extracteur de termes TermoStat Web 3.0 à partir d'un corpus de lignes de concordance KWIC (Key Word In Context) du terme qui désigne un concept à définir peut être une source efficace d'information définitionnelle. À cette fin, nous avons utilisé comme référence une liste de termes générés à partir d'un corpus anglais de définitions spécialisées du concept MAGMA (du domaine de l'environnement). Pour minimiser les interférences dues à la variation terminologique, les termes de la liste de référence ont été classés en fonction de la relation conceptuelle établie avec MAGMA. Ceci a été réalisé sans difficulté, car TermoStat Web 3.0 permet de consulter les concordances d’un terme dans la liste afin de déterminer la relation conceptuelle qu'il partage avec le concept à définir. Ensuite, la liste de référence a été comparée avec les listes de termes générées à partir de cinq corpus de concordances KWIC de magma (chaque corpus comporte des concordances d’une longueur différente) afin de déterminer la taille la plus appropriée du contexte en fonction du taux de précision et de rappel. Nos résultats, bien que préliminaires indiquent qu’un corps qu’un corpus de concordances KWIC de 250 caractères combiné avec la liste des termes générés à partir de celui-ci pourrait être une ressource utile dans la rédaction de définitions, soit comme complément d'un corpus de définitions, soit en remplaçant ce dernier s’il n’est pas disponible.

Abstract. This paper presents a pilot study that tested whether a list generated by the term extractor TermoStat Web 3.0 from a corpus of KWIC (Key Word in Context) concordance lines for the term designating a given concept to be defined can be an effective source of definitional information. For this purpose, a term list generated from an English corpus of specialized environmental definitions of the concept MAGMA was used as a reference. In order to minimize the interference from terminological variation, the terms in the reference list were categorized according to their conceptual relation with MAGMA. This was easily accomplished since TermoStat Web 3.0 allows the user to consult the concordances of a term on the list so as to determine their conceptual relation with the concept to define. Afterwards, the reference list was compared to the term lists generated from five corpora of KWIC concordance lines of magma with a different number of characters to determine the best length for the context, based on the precision and recall ratio. This also permitted us to derive preliminary conclusions regarding the usefulness of KWIC corpora. The results indicate that a 250-character KWIC corpus coupled with a term list generated from it could be a useful tool for the formulation of definitions, either as a complement to a definitions corpus or as a substitute for it when one is not available.

Mots-clés : corpus KWIC, extraction de termes, définition terminologique, extraction d’information définitionnelle

Keywords: KWIC corpus, term extraction, terminological definition, definitional information extraction
1. Introduction

This pilot study tested the performance of a special kind of term list as a source of definitional information. The list was generated by a term extractor from a corpus of KWIC (Key Word In Context) concordance lines of the term designating the concept to be defined. Such a list could be used to select the most relevant information for natural-language definitions. A semi-automatic extraction system of definitional information based on this list is also envisaged. The underlying hypothesis is that the high frequency of a lexical unit in a corpus of KWIC concordance lines of a term is an indication that the lexical unit could potentially be used in the definition of the term. In order to test this hypothesis, we conducted the pilot study described in the following sections.

2. Methods

We compared the precision and recall of a reference list obtained from a corpus of English definitions of the concept MAGMA with that of five analysis lists generated from corpora of KWIC concordance lines (KWIC corpora) of varying lengths for the term magma.

2.1. Reference list

The reference list was generated with the term extractor TermoStat Web 3.0 (Drouin 2003) from an English corpus of 133 specialized definitions of MAGMA (obtained from Metaglossary, Google Define, IATE, Termium, Le Grand Dictionnaire Terminologique, Oxford Reference, and Onelook). This corpus was fed into TermoStat in order to obtain the most frequent nouns, noun phrases, and adjectives. Only the terms with a frequency of over four occurrences were considered. So as to minimize interference from terminological variation, the terms in the reference list were categorized based on the most common conceptual proposition established with MAGMA in the concordances generated by TermoStat (see first and second columns of center table in poster).

It is true that any categorization inevitably has a certain degree of subjectivity. The configuration of our reference list is the result of certain choices. For example, temperature was included in the same category as hot because in the corpus it was mainly related to high temperatures and its frequency was relatively low. Moreover, liquid and fluid could have been categorized together with molten, but given their high frequency, it was decided that they should be in a separate category.

2.2. Analysis lists

Firstly, an environmental corpus of texts containing the lemma magma (359 occurrences) was compiled from English texts of the PANACEA Environment English monolingual corpus (http://catalog.elra.info/product_info.php?products_id=1184) and the LexiCon Research Group environmental corpus. Then, WordSmith Tools (Scott 2008) was used to generate KWIC concordance lines for magma with 100 characters before and 100 characters after the search term. The concordances were then saved in a single text file (100C corpus). The same procedure was followed to create KWIC corpora of 250 characters (250C corpus), 500 characters (500C corpus), and 750 characters (750C corpus). Concordances restricted to the sentences in which the lemma appeared were also generated (SC corpus). Finally, each corpus was fed into TermoStat in order to obtain the most frequent nouns, noun phrases, and adjectives. The 50 and 100 terms with the highest raw frequency were retained for comparison with the reference list. No more than 100 terms were retained because the objective was to test the usefulness of this list for a terminologist. Its length thus needed to be manageable.
2.3. Precision and recall calculation

We measured the precision ($P$) and recall ($R$) ratios using the following formulas as a means of comparison:

$$P = \frac{TP}{TP+FP}$$

$$R = \frac{TP}{TP+FN}$$

where:

- A true positive (TP) is defined as a term in the analysis list that matches any of the categories in the reference list. The result is expressed as a percentage. For a term to be considered a true positive, it only needs to be related to MAGMA as expressed in any of the conceptual propositions in the reference list. Derivatives, hypernyms, and hyponyms are also counted as true positives.
- A false positive (FP) is defined as a term in the analysis list that matches no category in the reference list. The result is expressed as a percentage.
- A false negative (FN) is defined as a category in the reference list that is not matched by any of the terms in the analysis list. The result is expressed as a percentage.

To combine both precision and recall ratios, we used the F2-measurement (Chinchor, 1992, 25), which gives twice the importance to recall as to precision. This formula was chosen because a term list generated from a KWIC corpus would eventually be used as a source of definitional information combined with other sources. Consequently, it is more important for all the relevant information to be retrieved despite a certain quantity of noise. The formula used was the following:

$$F2 = \frac{5 \cdot P \cdot R}{5 \cdot P + R}$$

3. Results

The results of the comparison of the reference list with the analysis lists are shown in the Results section of the poster. The 100-term 250C list performed the best (F2-M: 69.08 %). Moreover, its recall ratio was the highest (78.28 %). In terms of precision, the highest ratio corresponded to the 50-term 100C list. However, its recall ratio was almost 12 points below the 100-term 250C. Furthermore, as for the SC list, it obtained a lower F2 score compared to any of the KWIC lists. Finally, as can be observed, once the threshold of the 250-character context was exceeded, longer contexts caused both precision and recall to decrease. The third column of the center table in the poster shows the results of the comparison of the 100-term 250C.

4. Conclusions and future work

Although the scope of this pilot study was limited to the concept MAGMA and the lists generated only included nouns, nouns phrases, and adjectives, the results are encouraging. They indicate that a 250-character KWIC corpus coupled with a 100-term list generated from it could be a useful tool for the formulation of terminological definitions either as a complement to a definitions corpus or as a substitute for it when one is not available. In our opinion, given the objectives of this pilot study, the inevitable bias caused by the use of a reference list based on a manual classification does not invalidate the results.
This initial pilot study will subsequently be expanded to include new variables, such as the consideration of definienda other than physical entities (e.g. processes, artefacts, abstract entities, etc.), the inclusion of verbs and adverbs in the term lists, the use of corpora of different levels of specialization, and the generation of more KWIC corpora with different character counts. Moreover, the output of TermoStat will be compared with the output of other term extractors as well as that of a keyword generator against a blacklist. Once the hypothesis is definitively proven and the technique is refined, our ultimate objective is to combine it with the application of knowledge-pattern-based techniques (Pearson, 1998; Meyer, 2001; Malaisé et al., 2005; Marshman and L’Homme 2006; Auger and Barrière, 2008, inter alia) to create a system of semi-automatic definitional information extraction.

References


**Introduction**

This pilot study tested the performance of a special kind of term list as a source of definitional information. The list was generated by a term extractor from a corpus of KWIC (Key Word In Context) concordance lines of the term to be defined. Such a list could be used to test the most relevant information for natural-language definitions. A semi-automatic extraction system of definitional information based on this list is also envisaged.

The underlying hypothesis is that the high frequency of a lexical unit in a corpus of KWIC concordance lines of a term is an indication that the lexical unit could potentially be used in the definition of the term. In order to test this hypothesis, we conducted the pilot study described in the following sections.

**Methods**

We compared the precision and recall of a reference list obtained from a corpus of English definitions of the concept *magma* with that of five analysis lists generated from corpora of KWIC concordance lines (KWIC corpora) of varying lengths for the term *magma*.

**Reference list**

The reference list was generated with the term extractor TermoStat Web 3.0 (Drouin 2003) from an English corpus of 133 specialized definitions of *magma*. This corpus was fed into TermoStat in order to obtain the most frequent nouns, noun phrases, and adjectives. Only those with a frequency of over four occurrences were considered. So as to minimize interference from terminological variation, the terms in the reference list were then categorized based on the most common conceptual proposition established with MAGMA in the concordances generated by TermoStat (see first and second columns of center table).

**Analysis list**

An environmental corpus of texts containing the lemma *magma* (359 occurrences) was compiled from English texts of the PANACEA Environment English monolingual corpus and the LexiCon Research Group environmental corpus. Then, WordSmith Tools (Scott 2008) was used to generate KWIC concordance lines for *magma* with 100 characters before and 100 characters after the search term. The concordances were then saved in a single text file for each corpus.

Each corpus was fed into TermoStat in order to obtain the most frequent nouns, noun phrases, and adjectives. The 50 and 100 terms with the highest raw frequency were retained for comparison with the reference list. No more than 100 terms were retained because the objective was to test the usefulness of this list for a terminologist. Its length thus needed to be manageable.

**Precision and recall**

We measured the precision (P) and recall (R) ratios using the following formulas as a means of comparison:

\[ P = \frac{TP}{TP+FP} \]

\[ R = \frac{TP}{TP+FN} \]

A true positive (TP) is defined as a term in the analysis list that matches any of the categories in the reference list. The result is expressed as a percentage.

A false positive (FP) is defined as a term in the analysis list that matches no category in the reference list. The result is expressed as a percentage.

A false negative (FN) is defined as a category in the reference list that is not matched by any of the terms in the analysis list. The result is expressed as a percentage.

To combine both precision and recall ratios, we used the F2-measurement (Chinchor 1992, 25), which gives twice the importance to recall as to precision. The formula used was the following:

\[ F2 = \frac{2 \times P \times R}{P + R} \]

**Results**

The results of the comparison of the reference list with the analysis lists are shown in the following tables. For the sake of clarity, precision, recall, and F2-measurement are also represented in a graphic.

**Conclusions and future work**

Although the scope of this pilot study was limited, the results are encouraging. They indicate that a 250-character KWIC corpus coupled with a 100-term list generated from it could be a useful tool for the formulation of terminological definitions either as a complement to a definitions corpus or as a substitute for it when one is not available.

This initial pilot study will subsequently be expanded to include new variables and the output of TermoStat will be compared with the output of other term extractors as well as that of a keyword generator against a blacklist.

Once the hypothesis is definitively proven and the technique is refined, our ultimate objective is to improve precision and recall ratios. The third column of the experiment table shows the results of the comparison of the 100-term 250C corpus.

**References**


