How to exploit paralinguistic features to identify acronyms in texts?

Mathieu Roche

To cite this version:


HAL Id: lirmm-00974797
https://hal-lirmm.ccsd.cnrs.fr/lirmm-00974797

Submitted on 7 Apr 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
How to exploit paralinguistic features to identify acronyms in texts?

Mathieu Roche

UMR TETIS, Cirad, Irstea, AgroParisTech,
500, rue J.F. Breton, 34093 Montpellier Cedex 5, France
Mathieu.Roche@cirad.fr

LIRMM, CNRS, Univ. Montpellier 2,
161, rue Ada, 34000 Montpellier, France
Mathieu.Roche@lirmm.fr

Abstract
This paper addresses the issue of acronym dictionary building. The first step of the process identifies acronym/definition candidates, the second one selects candidates based on a letter alignment method. This approach has two advantages because it enables (1) to annotate documents, (2) to build specific dictionaries. More precisely, this paper discusses the use of a specific linguistic concept, the gloss, in order to identify candidates. The proposed method based on paralinguistic markers is independent of languages.

Keywords: text mining, acronym expansion

1. Introduction
Acronyms are numerous in specialized domain, e.g. biomedical and agronomy documents (Chang et al., 2002). An acronym is a set of characters corresponding to the first letters of a group of words, for instance, the acronym FAO is associated with the definition Food and Agriculture Organization. This paper summarizes a method to identify acronyms and expansions in documents. This automatic recognition enables to annotate these elements in texts. This work deals with the use of paralinguistic features in order to identify acronym/definition couples.

After the description of related work in the following section, Section 3, describes our approach based on 2 steps: Extraction of acronym/expansion candidates (Section 3.1.), Filtering of candidates (Section 3.2.). Finally, before Discussion and Conclusion sections, experiments of our approach are detailed in Section 4.

2. Related work
Among the several existing methods for acronym extraction in the literature, some significant work need to be mentioned. The acronym detection involves recognizing a character chain as an acronym and not as an unknown or misspelled word. Most acronym detecting methods rely on using specific linguistic markers.

Yates’ method (Yeates, 1999) involves the following steps: First, separating sentences by segments using specific markers (brackets, points) as frontiers. Then the acronym/definition couples are tested. The acronym/definition candidates are accepted if the acronym characters correspond to the first letters of the potential definitions words. The last step uses specific heuristics to select the relevant candidates. These heuristics rely on the fact that acronyms length is smaller than their expansion length, that they appear in upper case, and that long expansions of acronyms tend to use stop-words such as determiners, prepositions, suffixes and so forth. Therefore, the pair ”FAO/Food and Agriculture Organization” is valid according to these heuristics.

Other studies (Chang et al., 2002; Larkey et al., 2000) use similar methods based on the presence of markers associated with linguistic and/or statistical heuristics. In this context (Okazaki and Ananiadou, 2006) propose statistical measurements from the terminology extraction area. Okazaki and Ananiadou apply the C-value measure (Frantzi et al., 2000; Nenadic et al., 2003) initially used to extract terminology. It favors a candidate term that doesn’t appear often in a longer term. For instance, in a specialized corpus (i.e. Ophthalmology), the authors discovered that the term ”soft contact” was irrelevant, while the frequent and longer term ”soft contact lens” is relevant. An advantage of C-value measure is its independence from characters alignment (actually, a lot of acronyms/definitions are relevant while the letters are in a different order, e.g. ”AW / water activity”).

Other approaches based on supervised learning methods consist of selecting relevant expansions. In (Xu and Huang, 2007), the authors use SVM approaches (Support Vector Machines) with features based on acronym/expansion information (e.g. length, presence of special characters, context, etc). (Torii et al., 2007) present a comparative study of the main approaches (supervised learning methods, rules-based approaches) by combining domain-knowledge.

Larkey et al.’s method (Larkey et al., 2000) uses a search engine to enhance an initial corpus of Web pages useful for acronym detection. To do so, starting from a list of given
acronyms, queries are built and submitted to the AltaVista1 search engine. Query results are Web pages which URLs are explored, and possibly added to the corpus.

3. Acronym/expansion recognition

Our method of construction of acronym dictionaries is based on two steps detailed in the following subsections.

3.1. Step 1: Extraction of candidates

First, specific punctuation and character markers are taken into account in order to identify acronym/definition pairs (see Figure 1). In this paper, we investigate the extraction of candidates by exploiting the "glosses" of words and paralinguistic markers (i.e. brackets, punctuations, etc.) to detect acronym/definition candidates.

Glosses are spontaneous descriptions identifiable with specific markers (for example, called, i.e., and so forth). These ones highlight lexical semantic relationships, e.g. equivalence, specification of the meaning, nomination, hyponymy, hyperonymy.

The abstract pattern of glosses is given by the structure X marker Y₁, Y₂...Yₙ where X and Yᵢ can be acronyms and/or definitions. The identification and selection of glosses are based on the use of patterns and Web-mining approaches (Mela et al., 2012).

In this paper, we extract candidates based on the gloss markers "(" and ")":

- **Local Pattern 1** [X=acronym, Y₁=definition]: The first pattern detects Y₁ (definition), between "(" and ")" following the acronym (X). For example, the sentence "relation empirique entre l’indice de végétation NDVI (Normalized Difference Vegetation Index), mesuré au maximum ..." allows to extract X = NDVI and Y₁ = Normalized Difference Vegetation Index.

- **Local Pattern 2** [X=definition, Y₁=acronym]: The second pattern detects Y₁ (acronym), between "(" and ")" following the definition (X). The beginning of the definition is recognized with the first word of the phrase in upper case. For example, the sentence "$ \ldots la mesure Normalized Difference Vegetation Index (NDVI)" allows to extract X = Normalized Difference Vegetation Index and Y₁ = NDVI.

Note that these patterns are independent of languages because the method is based on paralinguistic markers (i.e., brackets in this work). This is very important when languages are mixed, for instance in specialized domains. The example of Figure 1 shows a definition in English (expansion of "NDVI") in an abstract written in French.

In this situation, we are 4 different cases of results:

- **Case 1**: the relevant definition is returned (like previous examples),
- **Case 2**: the extracted phrase contains the relevant definition (i.e. partially relevant, but too large),
- **Case 3**: the extracted phrase is a part of the relevant definition (i.e. partially relevant, but too specific),
- **Case 4**: the extracted phrase is irrelevant.

Both proposed patterns will be evaluated in Section 4. of this paper.

3.2. Step 2: Filtering of candidates

The second step aims at removing irrelevant acronym/definition pairs and deleting irrelevant word(s) from candidate definitions. For this process, we propose to align the acronym letters with the potential definition words, by mapping each acronym letter with the first character of each definition word, respecting the order of words. If the first letter of the candidate definition word can not be aligned with the acronym corresponding character, the following characters (of the word) are taken into account. For instance, this method allows to find that "Extraction Itérative de la Terminologie" is a possible definition of the French acronym EXIT.

4. Evaluation

This paper focuses on the study of a corpus of 2000 paper abstracts provided by Cirad2: French research centre working with developing countries to tackle international agricultural and development issues. Table 1 shows that better results are given with the second local pattern. But a lot of cases are partially relevant (i.e. ~ 40%), so we have to improve and enrich this pattern approach.

<table>
<thead>
<tr>
<th>Patterns</th>
<th>Local pattern 1</th>
<th>Local pattern 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of extracted definitions</td>
<td>78</td>
<td>64</td>
</tr>
<tr>
<td>Case 1 (relevant)</td>
<td>31 (39.7%)</td>
<td>28 (43.7%)</td>
</tr>
<tr>
<td>Case 2 (partially relevant)</td>
<td>3 (3.8%)</td>
<td>6 (9.3%)</td>
</tr>
<tr>
<td>Case 3 (partially relevant)</td>
<td>1 (1.3%)</td>
<td>18 (28.1%)</td>
</tr>
<tr>
<td>Case 4 (irrelevant)</td>
<td>43 (55.1%)</td>
<td>12 (18.7%)</td>
</tr>
</tbody>
</table>

Table 1: Evaluation of extracted definition with patterns.

The evaluation of the acronym/expansion extraction method is conducted on a corpus (general domain) having a reasonable size (7465 words). The experiments based on standard evaluation measures of data-mining domain highlight acceptable results (i.e. Precision: 66.7%, Recall:

---

1[www.altavista.com/](http://www.altavista.com/)
Un modèle de prévision de rendement de la canne à sucre basé sur des images satellites Spot : l’exemple de la Réunion

Auteurs

Boyer, Nathalie Todoroff, Pierre Paulin, Louis Bégué, Ağnès

Cet article présente une méthode de prévision de récolte de canne à sucre à partir d’images satellites à haute résolution spatiale Spot, adaptée à des zones de production de canne à sucre composées d’une multitude de petites parcelles. Cette étude est réalisée sur l’exemple de l’île de La Réunion. Elle est basée sur l’ajustement d’une relation empirique entre l’indice de végétation NDVI (Normalized Difference Vegetation Index), mesuré au maximum de développement du couvert végétal, et le rendement en biomasse fraîche à la récolte. Le modèle est appliqué en mode relatif pour estimer les variations de production d’une année à l’autre. Les variations de production estimées, ajoutées à la production mesurée de l’année précédente, permettent d’estimer la production de l’année à venir. La précision des prévisions de récolte ainsi réalisée est évaluée à partir des résultats de récolte des années 2008 et 2009. Les résultats sont très satisfaisants, avec une précision atteignant 6,6% à l’échelle des centres de récolte et 1,5% à l’échelle de l’île entière. Ils démontrent l’intérêt économique et technique d’une telle méthode dans le cas de zones de production disposant de peu d’informations sur les conditions de culture de la canne (état végétatif de nombreuses parcelles, météorologie contrastées). (Résumé de l’auteur)

Figure 1: Recognition of the couple \textit{NDVI / Normalized Difference Vegetation Index} in \textit{AGRITROP} database.

<table>
<thead>
<tr>
<th>Examples of extracted with Local pattern 1</th>
<th>Examples of extracted with Local pattern 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRPS</td>
<td>NonRibosomal Peptide Synthetase</td>
</tr>
<tr>
<td>VLE</td>
<td>Virtual Laboratory Environment</td>
</tr>
<tr>
<td>BMR</td>
<td>Bois Massif Reconstitué</td>
</tr>
<tr>
<td>ATPSM</td>
<td>Agricultural Trade Policy Simulation Model</td>
</tr>
<tr>
<td>ASA</td>
<td>Articulation du Semi-aride</td>
</tr>
<tr>
<td>CLF</td>
<td>Corynespora Leaf Fall</td>
</tr>
<tr>
<td>BASIC</td>
<td>Brésil, Afrique du Sud, Inde, Chine</td>
</tr>
<tr>
<td>CIAT</td>
<td>Centro international de agricultura tropical</td>
</tr>
<tr>
<td>BSV</td>
<td>Banana streak virus</td>
</tr>
<tr>
<td>ER</td>
<td>Ehrlichia ruminantium</td>
</tr>
<tr>
<td>CSSV</td>
<td>Cacao swollen shoot virus</td>
</tr>
<tr>
<td>MAE</td>
<td>Mesures agrienvironnementales</td>
</tr>
<tr>
<td>ACMV</td>
<td>African cassava mosaic virus</td>
</tr>
<tr>
<td>TYLCV</td>
<td>Tomato yellow leaf curl virus</td>
</tr>
</tbody>
</table>

Table 2: Examples of acronyms/definitions.

For instance, we can query a search engine with the acronym “BSV” and its possible definition to check on the Web if this association exists. This query should be a disjunction (i.e. OR operator) of the acronym and its possible definition returned with our process (i.e. Banana streak virus). This one returns a larger amount of documents. The conjunction of the acronym and the expansion (i.e. AND operator) enables to return a lower number of documents. But the returned documents are more relevant (i.e. the precision is improved).

In our case, we choose to consider the "hits" given by Google\footnote{http://www.google.fr/} on the examples of Table 2 (i.e. number of pages returned by the search engine based on conjunction). For instance, we have tested the query "BSV" AND "Banana streak virus" that returns 7580 pages\footnote{Queries performed on the 20th of March 2014}. All the results (i.e. hits) are given in Table 3. This table shows that hits have generally very high values, this allows us to automatically validate acronym/definition couples. Note that hits of irrelevant couples return lower values (for instance, with the couples "ETM"/"environ 5.000 m3.ha-1", "SIPSA"/"indicateurs, documents, cartes", and so on).

Moreover, we can integrate this kind of information in classical similarity measures, e.g. Dice measure \cite{Smadja1996}. Dice measure can be used to compute a sort of relationship between an acronym (i.e. \textit{acro}) and a definition (i.e. \textit{def}). In our context, Dice measure (formula (1)) is based on the number of Web pages given by search engines (i.e. hits).

\[
Web_{\text{dice}}(\text{acro}, \text{def}) = \frac{2 \times \text{hits}(\text{acro}, \text{def})}{\text{hits}(\text{acro}) + \text{hits}(\text{def})} \quad (1)
\]

80%, F-measure: 72.7\%) \cite{Roche2008}. We plan to apply the second step of the process (see Section 3.2.) with the pattern approach described in Section 3.1. on the Cirad corpus.

Note that our previous work \cite{Roche2008} uses more global patterns ; then a lot of noise is returned. The pattern approach described in this paper is more specific with better results in term of precision (\sim 40\% in this current work vs. 15\% in our previous work).

5. Discussion: Towards a Web-mining approach

In this section, we propose to integrate Web-mining measures in order to automatically validate results returned by our approach \cite{Turney2001, Mela2012}.
Table 3: Examples of acronym/definition and hits scores.

This measure returns the following result with the previous example:

\[
\text{WebDice}(\text{BSV, Banana streak virus}) = \frac{2 \times \text{hits}(\text{"BSV" AND "Banana streak virus"})}{\text{hits}(\text{"BSV"})+\text{hits}(\text{"Banana streak virus"})} = \frac{2 \times 7580}{2840000+15400} = 0.0053
\]

\text{WebDice} can be applied in order to rank couples (see Table 4). This enables to detect relevant acronym/definition pairs (i.e. couples with high \text{WebDice} values).

Table 4: Acronym/definition couples ranked with \text{WebDice}.

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Possible definition</th>
<th>\text{WebDice}</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATPSM</td>
<td>Agricultural Trade Policy Simulation Model</td>
<td>1.3014</td>
</tr>
<tr>
<td>NRPS</td>
<td>NonRibosomal Peptide Synthetase</td>
<td>230000</td>
</tr>
<tr>
<td>VLE</td>
<td>Virtual Laboratory Environment</td>
<td>36000</td>
</tr>
<tr>
<td>BMR</td>
<td>Bois Massif Reconstitué</td>
<td>9270</td>
</tr>
<tr>
<td>CIAT</td>
<td>Centro internacional de agricultura tropical</td>
<td>27700</td>
</tr>
<tr>
<td>ASA</td>
<td>Articulation du Semi-ardé</td>
<td>663</td>
</tr>
<tr>
<td>CLF</td>
<td>Corynespora Leaf Fall</td>
<td>22800</td>
</tr>
<tr>
<td>BASIC</td>
<td>Brésil, Afrique du Sud, Inde, Chine</td>
<td>21100</td>
</tr>
<tr>
<td>CSSV</td>
<td>Cacao swollen shoot virus</td>
<td>75000</td>
</tr>
<tr>
<td>MAE</td>
<td>Mesures agri-environnementales</td>
<td>951</td>
</tr>
<tr>
<td>ACMV</td>
<td>African cassava mosaic virus</td>
<td>90200</td>
</tr>
<tr>
<td>TYLCV</td>
<td>Tomato yellow leaf curl virus</td>
<td>354000</td>
</tr>
</tbody>
</table>

Moreover we plan to investigate a contrastive analysis of English/French corpora in order to give a new point of view of the phenomenon of spontaneous descriptions. A first study on aligned English/French texts reveals frequent regularities of glosses in a multilingual context. The alignment enables to improve the multilingual lexical acquisition of new words and their translations.

7. Acknowledgements
This work was supported in part by the French National Research Agency under JCJC program, grant ANR-12-JS02-01001, as well as by University Montpellier 2 and CNRS. The author thanks DIST (Scientific and Technical Information Service) for the acquisition of Cirad corpus.

8. References