Experiments for the Extraction of Qualia from Web in Italian: a Pattern-Based Approach

Fabio Celli
CLIC, university of Trento
fabio.celli@email.unitn.it

June 18, 2009

Abstract
This paper address the problem of automatically extracting qualia structures from the web via patterns in Italian. Qualia are treated as semantic relations, expressed by word patterns, between two entities. 400 examples of italian patterns expressing qualia are sampled and manually annotated in order to study the behaviour of the patterns and their suitability for qualia extraction. Features in the annotation schema include the position of the related entities with respect to the patterns, their informativeness and the minimum context window necessary to catch them. The results of this study are useful for designing experiments with word-space models.

1 Introduction and related work

Qualia structures, originally proposed by Pustejovsky [4], are simple and powerful semantic structures defining the characteristics of objects in the world. The task of extracting qualia from the web has already been tried in English by Cimiano and Wenderoth [2] among others, who adopted two different approaches: one patter-based, and the other one hit-based. For the task of qualia extraction in Italian here is adopted only the patter-based approach and the qualia are treated as semantic relations. Semantic information extraction via patterns is a well known approach in the NLP community since Hearst 1992 [3], who extracted “is-a” relations. In the next section are introduced the four qualia and the lexical patterns expressing them in Italian, then in section 3 are discussed the characteristics of those patterns and their suitability for qualia extraction.

2 Qualia structures and Experimental Settings

Qualia Structures The four qualia described in the Generative Lexicon framework [4] are: 1) the formal role, that describes the type or category from which an object inherits properties. 2) the constitutive role, that describes the parts of which the object is made of. 3) the telic role, that describes the purpose of an object. 4) the agentive role, that describes how or why an object is in the
Here the qualia are treated as semantic relations, expressed by patterns, between two linguistic entities (words), one that "has" (receives) and one that "is" (bears) the relation. For each of those qualia/semantic relation a set of lexico-syntactic patterns has been chosen by using the native speaker’s competence, qualia and patterns are reported in table 1:

In order to study the usage of those patterns in Italian 100 sentences for each pattern were extracted from ItWaC, [1], the largest general-purpose and web-based corpus available for Italian. No part-of-speech information were used in the queries in order to simulate the normal query condition in the web.

### Annotation Schema and Guidelines
The annotation schema includes 1) pattern, 2) related entity position with respect to the pattern (has-is, is-has, both to the right-has-is), 3) entity informativeness (nominal, verbal, sentential, pronominal, proper names), 4) bag-of-words (a real number between 1 and 30). Feature 3 is very useful to understand whether the related entities are easy to extract or not: in the case they are NPs (nominal) and VPs (verbal) they are easy to extract, in the case they are embedded in clauses (sentential context), or in the case they are just pronouns, they are clearly less informative and harder to detect and to extract.

The annotation guidelines were the following: 1) pattern detection and annotation, 2) related entity detection and bag-of-word count, 3) entity informativeness judgement annotation (based on the the fact if the entity were located in NPs, VPs, separated from the pattern by embedded clauses, or they were pronouns), 4) judgement and annotation about the position (left, right) and the function of the related entities (has-relation, is-related) with respect to the pattern. The position is annotated only if both the entities are located to the left or to the right of the pattern, otherwise it is sufficient to annotate the position with the functions (for example "is-has", "has-is").

### 3 Results and Discussion

Table 2 shows the frequency of the patterns. Of the four patterns extracted, "deriva d*” (derived from) and "è composto d*” (made of) showed two variants: "deriva da", "deriva di" and "è composto da", "è composto di". The patterns "deriva di” (the drift of), which have a different sense with respect to "deriva da” (derived from), were discarded, instead the examples "è composto di” (is made of) were counted henceforth as "è composto da” (is made by), since they have the same meaning. A closer look at the data reveals that the agentive pattern "deriva d*” has a bias in the fact that there are collocations expressing
come ad esempio

<table>
<thead>
<tr>
<th></th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>deriva da</td>
<td>96</td>
</tr>
<tr>
<td>deriva di</td>
<td>4</td>
</tr>
<tr>
<td>é composto da</td>
<td>91</td>
</tr>
<tr>
<td>é composto di</td>
<td>9</td>
</tr>
<tr>
<td>ha lo scopo di</td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 2: Pattern variants and frequencies**

derivation of words from other words or languages. Another problem shared by all patterns is the presence of metaphors like the following: "insicurezza che deriva dal crollo delle certezze" that would classify "crollo" and "certezza" in the agentive role of "insicurezza". Table 3 shows the distribution of the related elements with respect to the qualia pattern, and their part-of-speech. All the

<table>
<thead>
<tr>
<th></th>
<th>nm</th>
<th>np</th>
<th>ns</th>
<th>nv</th>
<th>pn</th>
<th>pv</th>
<th>sn</th>
<th>ss</th>
<th>sv</th>
<th>vn</th>
<th>vv</th>
<th>xx</th>
</tr>
</thead>
<tbody>
<tr>
<td>formal</td>
<td>68</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>10</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>agentive</td>
<td>69</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>7</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>constitutive</td>
<td>72</td>
<td>0</td>
<td>13</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>7</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>telic</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>75</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>21</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3: Qualia and part-of-speech of left and right related elements:**
n=nominal, v=verbal, p=pronominal, s=sentential, x=not related

proper names, annotated in the previous phase, were considered here as nouns in order to reduce sparseness. The values in boldface in table 3 reveal that, apart the telic role pattern which has a strong tendency to occur within a noun or noun phrase as the first element and a verb or verbal phrase as the second one, in general qualia patterns tend to occur within nouns or noun phrases. The values in italic show that for the telic and formal roles the related element to the left of the pattern is harder to extract because it tends to occur in long or embedded sentences, the opposite is true for the constitutive pattern, which has more elements occurring in sentential contexts to the right. The agentive pattern instead seems to have sentential contexts both to the left and to the right, and it shows also the highest frequency of pronominal patterns (with the personoun to the left). Interestingly, almost all examples of all the qualia patterns show a position of the related entities of the type "has-is". It means that the element to the left of the pattern receives the relation and the one to the right is the relation bearer. For example in "la legittimita deriva da una decisione del parlamento" (The legitimacy derives from a parliament’s decision) the NP decisione del parlamento is the agentive role of legittimita, which receives the relation.

The bag-of word within which are found the related elements are reported in figure 1. It shows clearly that a window size of 5 words left and 5 right, is the best one for the extraction of related items via patterns.
conclusions and future work

In this paper is presented a preliminary exploration of the behavior of Italian patterns for the automated extraction of qualia structures from the web. It was found that the formal, agentive and costitutive qualia patterns lead to the extraction of nouns and proper names, and that the telic qualia pattern leads to the extraction of nouns and verbs instead; for all the qualia the relation configuration is of the type has-is. It was also found that the best bag-of-word for the extraction is a window size of 5 words, that can be reduced in the case of the agentive and telic roles, and increased in the case of the constitutive. These findings can be exploited for further experiments using word space models.

References


