Knowledge Extraction and Multidimensionality in the Environmental Domain

Pilar León Araúz and Arianne Reimerink

Keywords: knowledge extraction, multidimensionality, corpus pattern analysis, knowledge pattern

1 Introduction

The systematization of knowledge extraction techniques would make knowledge representation a more empirical and efficient process. But first of all, certain selection criteria must be defined by manually identifying what kind of information is useful or not, why it is so and how it is structured, so that automatic processes can be improved with human implicit encyclopaedic knowledge. Knowledge patterns (KPs) have been widely explored and are considered as one of the most reliable methods in knowledge extraction (Cimiano and Staab 2005; Marshman and L’Homme 2006). However, many of them are domain-specific and cannot always be reused. Moreover, they often convey different conceptual relations and are therefore polysemic structures. The development of pattern-based constraints can help to disambiguate them and at the same time avoid conceptual noise. Nevertheless, another factor that must be taken into account in knowledge representation is contextual multidimensionality. A natural consequence of multidimensionality is conceptual dynamism, especially prevalent in process-oriented domains, such as the ENVIRONMENT. Even within the same specialized domain, concepts are involved in very different frames or situations and, consequently, they may experience a change in their relational behaviour that must somehow be represented. In the following sections we show how Corpus Pattern Analysis can tackle both dynamism and knowledge patterns in conceptual concordances.

2 Corpus Pattern Analysis

Corpus Pattern Analysis (CPA) (Pustejovsky, Hanks and Rumshisky 2004; Hanks and Pustejovsky 2005) investigates syntagmatic criteria for distinguishing different meanings of a polysemous predicate. The procedure consists of three components: (1) the manual discovery of selection context patterns for specific verbs; (2) the automatic recognition of instances of

---

1 This research has been supported by project FFI2008-06080-C03-01/FILO, from the Spanish Ministry of Science and Innovation.
2 Our research on environment is hosted in http://manila.ugr.es/visual/
the identified patterns; (3) the automatic acquisition of patterns for unanalyzed cases (Rumshisky et al. 2006: 329).

We apply the CPA approach in a slightly different way. On the one hand, the extraction of concordances starts with a direct search of specialized terms. After that, they are classified according to the dimensions they show (material, function, location, etc.), where different knowledge patterns can be associated to different conceptual relations (made_of, has_function, located_at, etc.). Then we analyse multidimensionality according to context, since conceptual dimensions are sometimes restricted to specific situations.

In this sense, Rumshisky et al. (2006) indicate several ways in which different context dimensions expressed in the selection context patterns can affect the semantic interpretation of a predicate. According to them, the most frequent source of meaning differentiation of verbs lies in contrasting the argument types filling each argument slot (Rumshisky et al. 2006:330). However, in our research CPA is applied to identify how context can affect the relational behaviour of concepts, mainly noun entities. In our experience, the most frequent source of context differentiation of a concept’s behaviour lies in contrasting the specific values filling each dimension. This does not mean that concepts are regarded as polysemic structures (this could only be the case of terms), but rather that the activation of different contexts makes certain conceptual relations (or dimension values) incompatible. In Tables 1 and 2 both approaches are shown.

Table 1: Concordances of abandon in three different senses according to the CPA approach

<table>
<thead>
<tr>
<th>Meaning differentiation of the verb abandon</th>
</tr>
</thead>
<tbody>
<tr>
<td>[[Human</td>
</tr>
<tr>
<td>[[Human</td>
</tr>
</tbody>
</table>

Concordances

did not agree that Britain could or should abandon development, either for itself or for the
After discreet soundings, they prudently abandoned the idea, which would have involved a

| [[Human | Institution]] abandon [[Attitude]] |
|----------------------------------------|
| [[Human | Institution]] ceases to have [[Attitude]] |

Concordances

had previously attacked, and in 1931 he abandoned the principle of free trade, which he had
s hope for a future society, which would abandon the obsession with private property, but

| [[Human 1 | Animal 1]] abandon [[Human 2 | Animal 2]] (to [[Anything = Bad]]) |
|-------------------------------------------------|
| [[Human 1 | Animal 1]] goes away from and ceases to care for or look after [[Human 2 | Animal 2]], with the result that [[Anything=Bad]] may get them or happen to them |

Concordances

3 http://deb.fi.muni.cz/pdev/?action=patterns&id=abandon
to join sects that make their adherents abandon their families, give up their money and
t ransferred by thousands of animals being abandoned or buried in the countryside, according

Table 2: Concordances of WATER in three different domains according to our corpus analysis approach

<table>
<thead>
<tr>
<th>Context differentiation of the concept WATER</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGRICULTURE</td>
</tr>
<tr>
<td>[water + (functional pattern: available for…) + agriculture-related concept]</td>
</tr>
<tr>
<td>supplies in the area. In areas where some water is available for irrigation shrubs planted on</td>
</tr>
<tr>
<td>WATER TREATMENT/SUPPLY</td>
</tr>
<tr>
<td>[water + (functional pattern: for use by…) + water treatment/supply-related concept]</td>
</tr>
<tr>
<td>accompanying increased demand of fresh water for use by a growing population. One of the</td>
</tr>
<tr>
<td>GEOLOGY</td>
</tr>
<tr>
<td>[water + (causative pattern: caused…) + geology-related concept]</td>
</tr>
<tr>
<td>s vegetation to hold soil in place, flowing water caused even more erosion and nutrient loss, which</td>
</tr>
<tr>
<td>ley has been limited and linear erosion by water has been a major geomorphic process in the all</td>
</tr>
</tbody>
</table>

According to our corpus analysis approach for noun specialized concepts, different collocational values activate diverse situations and referents that do not usually coincide, if ever, in time or space. In Table 2, WATER is still the same concept in all three examples, but the conceptual propositions in which it participates are context-dependent. In a more natural context, WATER is usually regarded as an agent, which explains why it is prototypically linked through causative patterns. However, in anthropic contexts, WATER is more generally regarded as an artifact or a patient, which is often illustrated through functional patterns.

On the other hand, a pattern-based search is also conducted in order to find new relations among other concepts. Since many KPs are codified by polysemic verbs, a more classical CPA approach can be applied to search for a “real” meaning differentiation. For instance, it is evident that form does not have the same meaning in the following sentences: clouds are formed of water and clouds are formed by condensation. In this case prepositions play an important role, but also do argument fillers. The second arguments of each sentence (water and condensation) have a different conceptual nature (entity and process) and imply different relations (made_of and result_of). Our approach differs in the type of features we use to disambiguate them. Whereas in classical CPA arguments are all classified as semantic categories, we also add other types of information, such as morpho-syntactic data. However, not all verbs used as KPs are polysemic structures, such as range or constitute. In these cases, only argument fillers and their specificity inform us about context, dimensions and relation types, which also avoid conceptual noise.

3 Conceptual dimensions

In the first step, apart from manually extracting all dimensions related to candidate terms, we found that conceptual dimensions do not have the same status or relevance, and they have different implications depending on their representational level and specific contexts. We
propose a three-level classification: prototypical dimensions, additional dimensions and
dynamic dimensions (León Araúz 2009).

3.1 Static knowledge: prototypical and additional dimensions
Prototypical dimensions are the most frequent dimensions found in the corpus. They match
correspondences of GROIN into a classification (Leon Araúz 2009).

Table 3: Conceptual dimensions found in definitions of GROIN

<table>
<thead>
<tr>
<th>Concept 1</th>
<th>Dimension</th>
<th>Concept 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Groin</td>
<td>Material</td>
<td>wood; stone, concrete, timber, rock; sheet-pile, rubble-mound</td>
</tr>
<tr>
<td>Groin</td>
<td>Location</td>
<td>coastline; shore; seaward; shore, water, sea, river; shore; seawall, backshore, foreshore; backshore, water body</td>
</tr>
<tr>
<td>Groin</td>
<td>Function</td>
<td>prevent erosion and longshore drift; control erosion; stop the flow of beach material moved by longshore drift; prevent a beach from washing away; protect the shore from erosion; reduce longshore currents; trap littoral drift, retard erosion of the shore</td>
</tr>
</tbody>
</table>

Most concordances showed the same information, through different KPs like made of,
constructed of, built out from, located along and function by.

Figure 1: Knowledge patterns and conceptual dimensions in concordances of GROIN

However, concepts can also be related in many other ways in the real world. Additional
dimensions add new knowledge in two ways: (1) new conceptual dimensions may arise, (2) or
conceptual dimensions remain the same, but they show different values from those found in
the definition, some of them having a different degree of focalization compared to definitional
ones, and some others belonging to diverse hierarchical paradigms.
In Figure 2, the material, location and function dimensions are still associated to GROIN. Nevertheless, GROIN appears related to more specific concepts: SAND BAGS, SURF ZONE, ANCHOR, etc. In addition, new dimensions like part and cause widen its relational power with new concepts like HEAD, LEE, ACCRETION and TURBIDITY. Additional dimensions are less relevant than prototypical dimensions and their inclusion in a knowledge base depends on its level of specialization.

The extraction of both prototypical and additional dimensions only relies on the reliability of KPs, which will be dealt with in section 4. However, dynamic dimensions need a CPA approach, since any specialized domain contains sub-domains in which conceptual dimensions become more or less salient depending on the activation of specific contexts.

3.2 Dynamic dimensions: contextual knowledge

Dynamic dimensions reflect a contextual multidimensionality derived from the situated nature of concepts. They are determined by context and occur especially in the case of concepts with a low degree of specificity. We call them versatile concepts because they are involved in a myriad of events. However, even if they appear in different contexts, they are not always related to the same concepts or through the same relations, especially in such a wide domain as the environment, where entities are subject to constant transformation.

For example, the concept SAND is generally (or prototypically) defined as a kind of sediment located in the sea, rivers or soil layers. A specialized definition could even include some information about its grain size. However, in real texts, SAND activates many other relations. In a more general domain, such as GEOLOGY, the concept is linked to others through: type, as a kind of SEDIMENT; attribute, related to grain size as a classification parameter; and material, linking the concept to the natural elements of which it forms part (VALLEY, SOIL, AQUIFER, DESERT, etc).
These dimensions could be considered additional ones. However, some of them are not compatible in other contexts. In a COASTAL PROCESS domain, salient dimensions become: material, although natural elements are restricted to coastal ones (SAND BARRIER, SAND BERM, SAND SPIT, BEACH, etc.); and patient, where the concept is involved in certain natural processes (WAVE ACTION, STORMS, LONGSHORE CURRENT, DEPOSITION, etc.).

If context is again restricted to the COASTAL DEFENCE domain, dimensions are still the same, but values are focalized to artificial elements (FENCE, BERM, DIKE) or processes (TRAPPING, PUMPING, DUMPING). Furthermore, there is a new dimension, highlighting the functional nature of the concept in this specific context (SAND protects DUNE-BLUFFS, SAND BODY is used for BEACH NOURISHMENT, etc.).

These three domains form a hierarchy (GEOLOGY → COASTAL PROCESS → COASTAL DEFENCE), but in a completely different domain, changes are more remarkable.
In the WATER TREATMENT domain, a new dimension is found, where SAND is linked to a particular instrument used in water treatment plants. The functional dimension now has a different value (FILTRATION) and patient and material are no longer representative conceptual dimensions.

Context is a dynamic construct that triggers or restricts knowledge (León Araúz et al. 2009). Moreover, Yeh and Barsalou (2006) claim that when situations are incorporated into a cognitive task, processing becomes more tractable than when situations are ignored, and the same can be applied to knowledge acquisition processes. As a result, a more believable representational system should account for re-conceptualization according to the situated nature of concepts, which would enhance knowledge acquisition.

In our research, we are developing context-based conceptual networks by dividing the environmental field in different contextual domains. Nevertheless, first we need to know which concepts are activated in each situation and how to extract this information. This is where CPA can help to accomplish our aim.

According to CPA, patterns expressing contextual information delimit the lexical senses of words. In our approach, patterns expressing contextual information delimit conceptual dimensions, which at the same time delimit domain membership.

Two procedures must be applied. First of all, if a concept only activates a particular dimension in just one domain, the identification of any KP linking the concept to that dimension will be enough to associate the concept to a concrete domain. In the case of SAND, if a KP expresses the instrument dimension, the concept will be automatically assigned to the WATER TREATMENT domain.

Nevertheless, domain disambiguation is not that easy when different domains can activate the same dimensions and, as a result, KPs are usually the same. For example, if SAND is found next to KPs like consist of, comprising, formed from, containing, or composed of, the material dimension ascribes concepts to three possible domains: GEOLOGY, COASTAL PROCESSES and COASTAL DEFENCE. Sometimes, certain KPs express a particular dimension in just one domain, such as made of, where SAND is always related to COASTAL DEFENCE because the pattern needs the activation of an artificial concept.
However, most of the time, this is not the case, and domain disambiguation requires a second step based on the kind of values associated to each dimension. At this stage, semantic annotation seems the only way to differentiate domain membership. As mentioned above, in the GEOLOGY domain materials must be natural elements found in nature, in the COASTAL PROCESS domain, materials are restricted to those found in the coastal area, and in the COASTAL DEFENCE domain materials are no longer natural elements. Consequently, annotation should be concept-oriented, differentiating all concepts in the hierarchy and assigning each of them to particular contexts. In this way, when dimensions are common to several domains, dimensional values are the only thing that can guide the knowledge extraction process. Semantic annotation has not been implemented yet, since first of all we need to know which kind of information (e.g. conceptual, semantic, morpho-syntactic) must be included, which is also the case for KPs.

4 Knowledge patterns

Many KPs can be found in a manual identification process. However, automatic extraction needs a certain level of reliability to be effective. In table 2 we show some of the most reliable patterns for seven conceptual relations in our specialized domain.

Table 3: Knowledge patterns and their conceptual relations

<table>
<thead>
<tr>
<th>Conceptual relation</th>
<th>Knowledge pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>Is_a</td>
<td>such as, rang* from, includ*</td>
</tr>
<tr>
<td>Part_of</td>
<td>includ*, consist* of, formed by/of</td>
</tr>
<tr>
<td>Made_of</td>
<td>consist* of, built of/from, constructed of, formed by/of/from</td>
</tr>
<tr>
<td>Located_at</td>
<td>form* in/at/on, found in/at/on, tak* place in/at, located in/at</td>
</tr>
<tr>
<td>Result_of</td>
<td>caused by, leading to, derived from, formed when/by/from</td>
</tr>
<tr>
<td>Has_function</td>
<td>designed for/to, built to/for, purpose is to, used to/from</td>
</tr>
<tr>
<td>Effected_by</td>
<td>carried out with, by using</td>
</tr>
</tbody>
</table>

We have applied CPA to certain general patterns. For example, despite all the information that form* helps to extract, it would have been a non-reliable pattern if no prepositions or verb tenses were included as part of the pattern. Thus, when form* is followed by in, at or on, it only expresses the relation located_at.

Domain-specific patterns are generally more reliable, such as built of/from and constructed of, which have 100 per cent reliability. But most of them are general language patterns that can be applied to many other domains. On the other hand, even reliable ones show a certain degree of conceptual noise. A way of solving this problem could be to search for candidate terms of interest occurring with patterns indicating relations that may be useful for describing the concepts they denote, but that would rule out unidentified terms.

Semantic annotation seems again a plausible way to exclude useless information. For example, in the KP consist* of, a higher level of reliability would be achieved if automatic searches could be restricted to part of speech nouns, since parts or materials will never be
anything other than nouns. In this way, sentences like “the process consists of building a groyne 97m long” would be omitted. On the other hand, if concept types were also annotated in the corpus, it would be easier to deal with KPs like carried out with. In this case, only concepts sharing the category membership of instrument would be selected, avoiding sentences like “two test series were carried out with the following parameters”.

This can also be applied to KPs expressing several conceptual relations, as in the case of domain disambiguation. Certain criteria could be elaborated for patterns like includ* (type_of and part_of), consist of (part_of and made_of) or formed by (part_of, made_of and result_of). Nevertheless, before annotating any corpus, it is necessary to discover certain rule constraints according to each KP’s specific needs, whether they cause conceptual noise or they are polysemic structures. In the next sections we will deal with the KPs rang* from and formed by.

3.1 KP rang* from

One of the most reliable knowledge patterns that activates the relation is_a in the environmental domain is rang* from. In our corpus there are 1381 occurrences of this pattern. Still, noise appears when studying the concordances with this knowledge pattern. When rang* from is followed by a number (see Figure 7), the relation that is expressed is always of magnitude. This is interesting information for the description of prototypical and additional dimensions of some types of concepts, but not if our aim is to analyze the conceptual relation IS_A. If we omit all these concordances, there are 356 concordances left. The same goes for the combination with a number written in full, or an adverb or an article and a number. Measurable concepts such as GROWTH TIMES and RECHARGE EVENTS are defined by certain time and amount combinations, which are additional dimensions. Therefore, if our aim is to analyze the IS_A relation, there should also be restrictions on words expressing duration such as minute, hour, month, day, week, etc. (see Figure 8).

Figure 7: Noise in the KP rang* from: numbers

If these concordances are omitted, there are only 121 concordances left. The knowledge pattern rang* from has proven to be a very reliable pattern if the above-mentioned restrictions...
are taken into account. It is not only very informative for the extraction of hyponymic relations, but also for relations of coordination, as the expression requires at least two coordinate concepts.

From the concordances of figure 9, the following information can be extracted: FINE SAND and PEBBLE is_a BEACH SEDIMENTS; LONG BEACH and STRAIGHT BEACH is_a BEACH; SAND, MUD, CLAY and BOULDER is_a SEDIMENT; PHAGOTROPHIC INGESTION is_a FEEDING STRATEGY.

### 3.2 KP formed by

As in the case of *rang* from, certain constraints must also be defined for the KP formed by in order to avoid noise. First of all, concordances where the KP is followed by a figure or a time expression (formed by then, formed by 2.5 billion years ago) must be omitted.

Moreover, it is also possible that the KP links a specialized concept with a too general one, both to the right and the left of the pattern.

Finally, a significant problem common to all KPs, as in *rang* from, is the fact that the first term is so far from the pattern, that it cannot even be manually extracted.
Noise shown in figures 11, 12 and 13 could be partly solved if a candidate term extractor could be implemented at the same time. This way, terms which did not fall under the consideration of specialized terms would be excluded from the extraction process from the beginning. Anaphora would still be a problem, but the validation of conceptual propositions would be at least more efficient. However, the main problem of the KP formed by is polysemy. Concordances in figure 14 show the way formed by works in the three different dimensions it can express, although the result dimension prevails over part and material.

The disambiguation of this polysemic KP requires different steps. If the KP is followed by a verb, it is definitely related to the result dimension. Instead, if the KP is followed by a noun, it can be linked to any of the three dimensions. Then the difference lies in two factors: if the noun is a process concept type, the concept still falls into the result dimension; if the noun is an object concept type, the dimension can be either part or material, but if the noun is uncountable, it will always refer to the material dimension, whereas countable nouns will always link wholes with parts.

In these concordances the word span among concepts and patterns is still quite wide, although there are no cases of anaphora. For formed by, a candidate term extraction process should also be running at the same time, where a 10 word maximum span would be set on the left side of the KP and a 4 word maximum span on the right.

5 Conclusions and future work

In order to construct a knowledge resource based on the real behaviour of concepts in all the possible contexts of a domain, all the dimensions of multidimensional concepts must be taken into account. Therefore, apart from providing a definition which conveys their prototypical
features, additional and dynamic dimensions must also be made explicit. In our research, we use the different dimensions of concepts to restrict the conceptual information shown to the end user. Instead of showing all the dimensions together, users can restrict their searches according to context and consequently avoid overinformation.

On the other hand, we have shown how constraints can be defined for KPs to facilitate knowledge extraction. This makes the terminographer’s job easier and provides the end user with more explicitly related information. Needless to say, there is still much work to be done before achieving an automatic knowledge extraction system. For example, even if all KPs were tightly constrained, we are still analysing terms, and synonymous values may seem different concepts. That could be solved if knowledge representation and extraction processes were complementary. In this way, ontological systems could inform the extraction system and semantic annotation could be based on an already defined hierarchy. In any case, manual validation would still be necessary. Once we extract all the information we need to avoid noise and disambiguate patterns and dimensions, we are planning to manually annotate a significant part of the corpus to use it as a gold standard. In this way, automatic procedures could next be implemented and slowly refined.

References