Knowledge Extraction on Multidimensional Concepts: Corpus Pattern Analysis (CPA) and Concordances[1]

Pilar León Araúz, Arianne Reimerink and Pamela Faber[2]

[2] Department of Translation and Interpreting, University of Granada

Abstract: Multidimensionality of concepts in multidisciplinary domains is a problem terminographers have to deal with. We apply Corpus Pattern Analysis (CPA; Pustejovsky, Hanks, & Rumshisky, 2004) to extract conceptual dimensions according to context. The dynamic nature of these concepts is exemplified with the case study of SAND. On the other hand, knowledge patterns (KPs) 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, which would be a first step towards the systematization of automatic knowledge extraction. Two KPs are analyzed in detail: *rang* from*, which conveys the conceptual relation *is_a*, and the polysemic KP *formed by*.

Keywords: Knowledge extraction, Multidimensionality, Corpus Pattern Analysis, Knowledge pattern.

Résumé: La multidimensionalité conceptuelle dans les domaines multidisciplinaires est un problème auquel les terminographes doivent faire face. On a appliqué le *Corpus Pattern Analysis* (CPA ; Pustejovsky et al., 2004) afin d'extraire les dimensions conceptuelles selon le contexte. La nature dynamique qui caractérise les concepts est illustrée avec l'exemple de SAND. Par ailleurs, les patrons de connaissance (KPs) expriment très souvent de différentes relations conceptuelles, étant ainsi des structures polysémiques. Dans ce sens, le développement de contraintes basées sur les KPs peut être une bonne approche pour contourner le problème de la polysémie et en même temps éviter le bruit conceptuel dans les concordances. Cela représenterait un premier pas vers la systématisation de l'extraction automatique de connaissances. Dans cet article, on présente l'approche suivi dans l'analyse de deux KPs: *rang* from*, exprimant la relation *is_a*, et *formed by*, un des KPs les plus polysémiques.

Mots clés: Extraction de connaissances, Multidimensionalité, Corpus Pattern Analysis, Patron de connaissance.

[1] This research has been supported by project FFI2008-06080-C03-01/FILO, from the Spanish Ministry of Science and Innovation.

1 Introduction

According to Kageura (1997: 120) the characteristics of a concept are frequently specified from different points of view or facets (function, material, shape...) and the set of characteristics that constitutes a concept is normally multidimensional. Moreover, when concepts are represented in different contexts (i.e. specialized subdomains) certain facets or dimensions become more or less salient. Corpus linguistics provides the clues to distinguish which dimensions are activated in each case and a sound methodology must be applied to study this contextual multidimensionality in a consistent way.

Corpus Pattern Analysis (CPA; Pustejovsky, Hanks, & Rumshisky, 2004; Hanks & 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 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. We analyse concordances starting with a direct search of specialized terms. After that, they are classified according to the dimensions they show, where different knowledge patterns (KPs; Barrière, 2004) can be associated to different conceptual relations. Then we analyse multidimensionality according to context. 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* (idem, 330). However, here CPA is applied to identify how context can affect the relational behaviour of a concept. In our experience, *the most frequent source of context differentiation of a concept's behaviour lies in contrasting the specific values filling each dimension*.

On the other hand, a pattern-based search is conducted in order to find new relations among other concepts. The problem is that knowledge patterns are very often polysemic structures that convey different conceptual relations. As a result, the development of pattern-based constraints can help to disambiguate them and at the same time avoid conceptual noise.

2 Conceptual dimensions

Contextual multidimensionality is derived from the situated nature of concepts. It occurs 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 and they are not always related to the same concepts or through the same relations, especially in interdisciplinary domains.

For example, the concept SAND is generally (or prototypically) defined as a kind of sediment located in the sea, rivers or soil layers. However, in real texts, SAND activates many other dimensions. 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 is part (VALLEY, SOIL, AQUIFER, DESERT, etc):

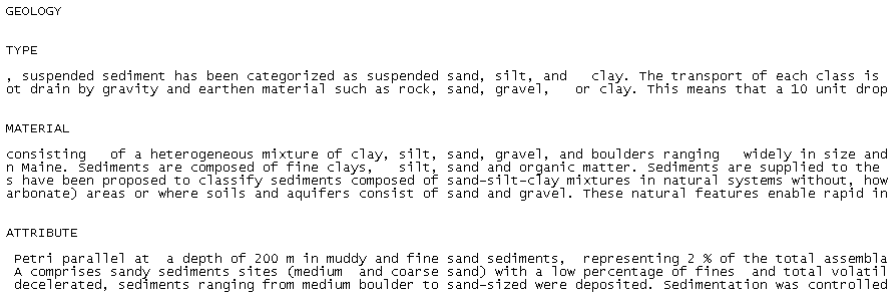


Fig. 1 – SAND in the GEOLOGY domain.

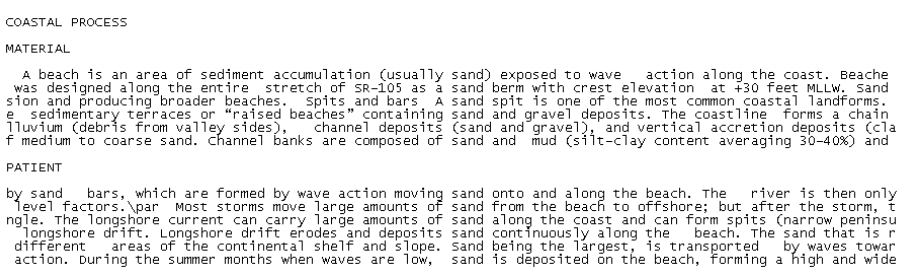


Fig. 2 – SAND in the COASTAL PROCESS domain.

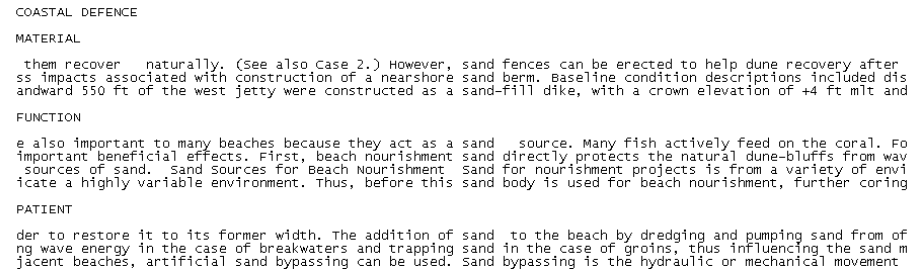
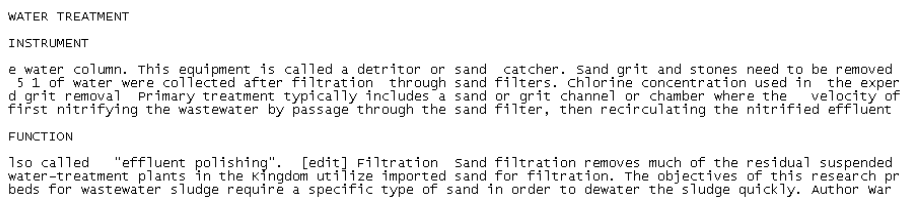


Fig. 3 – SAND in the COASTAL DEFENCE domain.

However, in a COASTAL PROCESS domain (see Fig. 2), salient dimensions become: *material*, although values (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). If context is again restricted to the COASTAL DEFENCE domain (see Fig. 3), 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:


WATER TREATMENT

INSTRUMENT

e water column. This equipment is called a detritor or sand catcher. Sand grit and stones need to be removed
5 1 of water were collected after filtration through sand filters. Chlorine concentration used in the exper
d grit removal Primary treatment typically includes a sand or grit channel or chamber where the velocity of
first nitrifying the wastewater by passage through the sand filter, then recirculating the nitrified effluent

FUNCTION

lso called "effluent polishing". [edit] Filtration Sand filtration removes much of the residual suspended
water-treatment plants in the Kingdom utilize imported sand for filtration. The objectives of this research pr
beds for wastewater sludge require a specific type of sand in order to dewater the sludge quickly. Author War

Fig. 4 – SAND in the WATER TREATMENT domain.

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.

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.

We are developing context-based conceptual networks by dividing the interdisciplinary field of the environment[2] in different contextual domains, but 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. In our approach, patterns expressing contextual information delimit conceptual dimensions, which at the same time delimit domain membership.

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

[2] http://manila.ugr.es/visual/

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.

3 Knowledge patterns

Many KPs can be found in the manual identification process. However, automatic extraction needs a certain level of reliability to be effective. In table 1 we show some of the most reliable patterns for seven conceptual relations in our specialized domain:

Table 1. Knowledge patterns and their conceptual relations.

Relation	**Knowledge pattern**
Is_a	such as, rang* from, includ*
Part_of	include*, consist* of, formed by/of
Made_of	consist* of, built of/from, constructed of, formed by/of/from
Located_at	form* in/at/on, found in/at/on, tak* place in/at, located in/at
Result_of	caused by, leading to, derived from, formed when/by/from
Has_function	designed for/to, built to/for, purpose is to, used to/for
Effected_by	carried out with, by using

Most patterns are general language expressions and can be applied to many other domains. Domain-specific patterns are generally more reliable, such as *built of/from* and *constructed of*. However, even reliable ones show a certain degree of conceptual noise and polysemy. Thus it is necessary to discover certain rule constraints according to each KP's specific needs. 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 KPs that activates the relation *is_a* in the environmental domain is *rang* from*. However, when *rang* from* is followed by a number (see Fig. 7), the relation that is expressed is always of magnitude. The same goes for the combination with a number written in full, or an adverb or an article and a number. For example, measurable concepts such as GROWTH TIMES and RECHARGE EVENTS are defined by certain time and amount combinations. 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 Fig. 5).

Furthermore, if the terms are too far away from the KP, it is very hard to extract useful information from the concordance (see Fig. 6). This noise could be partly solved with the implementation of a candidate term extractor. 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.

ditions. Temperatures in Italian rice fields typically range from 15 to 30 øC (Schütz et al., 1990). We chose
ved at different locations with aggregation timescales ranging from 15 min to 2 days. The theoretical and obse
a marked thermocline around 30 m depth; temperatures ranged from 15 to 17Â°C in the epilimnion and 7Â°C in
by land clearing, grubbing, grading, and construction range from 149,459 to 198,494 acres, or 25 to 33 perce
he same image analysis system as in the present study ranged from 15 to 90 mg C mâ€“3 from March to October
ow downgradient wells. Water table depth in the wells ranged from approximately 5 to 8 feet below the ground
eas that are stratified into 79 beach sampling units, ranging from ca. 3 to 37 km in length, and which are m
9]). Throughout the collected samples in 1999, the TDS ranged from a minimum of 5900 mg/l to a maximum of 14,2
cannot proceed much because the bedrock is at a depth ranging from about 100 to 120 m. However, several secto
ater aquifers of North Carolina, pesticide detections ranged from less than 1 to 34 percent of standards or
tandards or guidance levels. The remaining detections ranged from less than 1 to 50 percent of standards or
n plane and relatively steep beaches with growth times ranging from hours to one month. Less steep equilibrium

Fig. 5 – Noise in the KP *rang* from*: numbers and magnitude.

n, 1993; Mehrotra and Divyā, 1994). These studies have ranged from statistical models to conceptual rainfall-r
three-dimensional computer pictures of ocean animals, ranging from plankton to large fish. This new technolog
and among the organisms themselves. These interactions range from the formation of tiny plankton patches that
t to sea via rivers, streams and storm drains. Objects ranging from detergent bottles, hazardous medical waste
aquifers, estuaries and wetlands. The existing methods range from quick and simple assessments, where environm
a wide range of other survey and analytical techniques ranging from remote sensing to sampling of cetacean tis
on for relatively small projects. Applicable projects range from docks or riprap revetments in artificial la
The authors examine four different model formulations, ranging from simple deterministic optimization based on
y, these approaches have a number of problems, ranging from obvious visual impacts to the elimination
spectrum of biological and ecosystem organization, and ranging from selected biomarkers and tissue responses,
f biodiversity change that are occurring across scales ranging from species to ecosystems, rather than treatin
dies several kilometres wide and up to 15m thick. They range from massive to well-stratified and show large in

Fig. 6 – Inadequate concordances of the KP *rang* from*.

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:

sheries 1998). Marine waters The marine environment ranges from the intertidal zone to the vast waters of t
ia testudinum). Seagrasses normally occur in sediments ranging from sand to mud in relatively protected envir
occur in both freshwater and marine environments that range from pelagic to benthic and littoral to deep-sea
number of species with different feeding strategies, ranging from phagotrophic ingestion of particles to ph
ow-to-moderate wave energy-coasts with beach sediments ranging from fine sand to pebbles [Dally and Pope, 1986
rada (Tarragona) is composed of different beach-types, ranging from long, straight beaches to pocket beaches
er. where a variety of coarse material exists, ranging from pebbles to large boulders, one fin
three-dimensional computer pictures of ocean animals, ranging from plankton to large fish. This new technolog
ciomarine sediments (which is a mixture of sediment, ranging from clay to boulders). These were later rewo

Fig. 7 – Hyponymic and coordination concordances of the KP *rang* from*.

From figure 7, 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.

3.2 KP *formed by*

As in the case of *rang* from*, certain constraints must be applied in order to avoid noise. However, the main problem of the KP *formed by* is polysemy. Concordances in figure 8 show the way *formed by* works in the three different dimensions it can express, although the *result* dimension prevails over *part* and *material* (see Fig. 8).

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*.

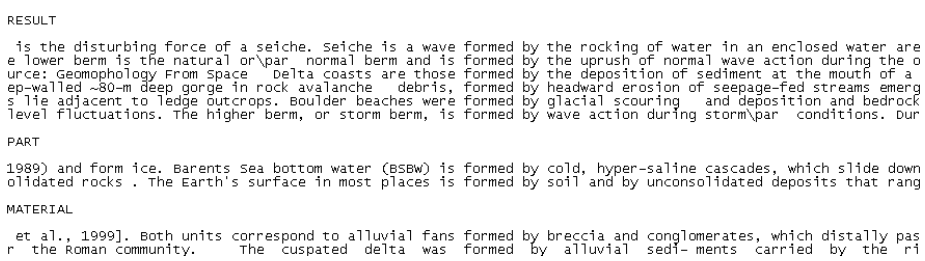


Fig. 8 – Conceptual dimensions expressed by KP *formed by*.

4 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. On the other hand, we have shown how constraints can be defined for KPs to facilitate knowledge extraction. 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 synonymic values may seem different concepts. That could be solved if knowledge representation and extraction processes were complementary. In this way, an ontological system 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.

References

HANKS, P. & PUSTEJOVSKY, J. (2005). A Pattern dictionary for natural language Processing. *Revue Française de linguistique appliquée* 10: 2, p. 63-82.

BARRIÈRE, C. 2004. "BUILDING A CONCEPT HIERARCHY FROM CORPUS ANALYSIS". *Terminology* 10: 2, 241-263.

PUSTEJOVSKY, J., P. HANKS, & RUMSHISKY, A. (2004). Automated induction of sense in context. *Proceedings of the 20th international conference on Computational Linguistics*. Geneva, p. 924-931.

RUMSHISKY, A., HANKS, P., HAVASI, C. & PUSTEJOVSKY, J. (2006). Constructing a Corpus-based Ontology using Model Bias. *Proceedings of FLAIRS 2006*, Melbourne Beach, Florida, 327-332.

YEH, W., & BARSALOU, L.W. (2006). The situated nature of concepts. *American Journal of Psychology*, 119, p. 349-384.