Knowledge-Rich Contexts Discovery

Barrière, Caroline

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Knowledge-Rich Contexts Discovery *

Barrière, C.
May 2004

* published at the Seventeenth Canadian Conference on Artificial Intelligence (AI'2004). London, Ontario, Canada. May 2004. NRC 48076.

Canada

Knowledge-rich Contexts Discovery

Caroline Barrière

Research Center for Language Technology, National Research Center of Canada,
Gatineau, Québec, Canada[1]
Caroline.Barriere@nrc-cnrc.gc.ca

Abstract. Within large corpora of texts, Knowledge-Rich Contexts (KRCs) are a subset of sentences containing information that would be valuable to a human for the construction of a knowledge base. The entry point to the discovery of KRCs is the automatic identification of Knowledge Patterns (KPs) which are indicative of semantic relations. Machine readable dictionary serves as our starting point for investigating the types of knowledge embodied in definitions and some associated KPs. We then move toward corpora analysis and discuss issues of generality/specificity as well as KPs efficiency. We suggest an expansion of the lexical-syntactic definitions of KPs to include a semantic dimension, and we briefly present a tool for knowledge acquisition, SeRT, which allows user such flexible definition of KPs for automatic discovery of KRCs.

1 Introduction

Texts, corpus of texts, are invaluable sources of information. But even if text has the quality of being rich and abundant, it has the default of being unstructured, sequential access, uneven in its "valuable information" rate, disorganized, ambiguous, sometimes redundant or even contradictory. This suits humans just fine when reading casually, but is a large obstacle to any attempt to automatic construction of domain model from text. If we settle for a slightly less ambitious goal of semi-automatic acquisition of information from texts, we still face the challenge of identifying the subset of sentences that contain valuable information.

The notion of value of information is quite subjective and is dependent on the task envisaged. For example, the construction of a concept dictionary [1] to help in question answering does not emphasize the same information as the construction of a Terminological Knowledge Base to help terminographers generate term definitions. As our research over the past years has been closer to the Computational Terminology community, we will adopt that view point in our investigation. Being aware of the large interest in knowledge discovery by the Information Extraction (IE) community, we will also include some comments relating to IE throughout the article.

Meyer [2] refers to those valuable sentences as Knowledge Rich Contexts (KRCs). Not only would a KRC contain a term of interest in a particular domain, but

[1] This research has been performed while the author was at the School of Information Technology and Engineering, University of Ottawa.

it would also contain a Knowledge Pattern (KP) showing how to link it to other terms of the domain. Knowledge Patterns (KPs) are the key to the discovery of Knowledge-Rich Contexts, so that the terminographer need not to look at thousands of sentences in an attempt to define a term, but can look at only the subset of sentences that contain the needed information.

We suggest, in this research, to deepen our understanding of Knowledge Patterns (KPs). Although they are commonly used toward knowledge discovery, although lists are made, although variations are searched for, they are not often discussed with regards to their definition and value. To achieve our goal, we decided to go back to the source and look into dictionary definitions for the types of knowledge embodied and their expression via KPs. This will be presented in Section 2.

In Section 3, we move from dictionary to real text corpora. Although research on dictionary can inspire and guide research on real text, it is not obvious it entirely reflects the type of knowledge necessary to understand a particular domain. Focusing on a specific domain of scuba-diving, looking at a 1M word corpus, we will address questions of generality/specificity of semantic relations and associated KPs.

In Section 4, in light of many examples of KPs presented in Section 2 & 3, we will suggest to expand the usual lexical-syntactic view of KPs to include a semantic dimension. As a KP is an entry point into a text, we want to be able to define it in a flexible way. This section will also explore the notion of productivity for KPs.

Section 5 briefly presents SeRT (Semantic Relations in Text) which allows for the definition of KPs and their use toward the discovery of Knowledge Rich Contexts.

Section 6 concludes.

2 Knowledge Patterns and Types of Knowledge

This section gives a possible organization of the type of information contained in dictionary definitions. This will serve for our discussion on corpora as we come back on the application-dependent notion of valuable information in Section 3.

The predominant relation found in a dictionary is the hyperonymy. Defining a word via its genus (superclass) dates back to Aristotle and persists today. The hyperonymy is part of the set of paradigmatic relations, which includes also synonyms and antonyms. We refer to this paradigmatic view of the conceptual model as **static knowledge**.

Table 1. Paradigmatic relations and their KPs

Semantic Relation	Example from AHFD
Opposite	Back **is the opposite of** front.
Synonymy	Automobile **is another word for** car.
	Earth **means** dirt.
Hyperonymy	An acorn **is a** nut that grows into an oak tree.
	An apple **is a kind of** fruit.
	Dogs, cats**,** birds**, and** insects **are all** animals.

This type of knowledge is very much present in dictionary definitions. Table 1 presents some examples of these paradigmatic relations with some typical KRCs taken from the American Heritage First Dictionary (AHFD).[2]

In fact, the static knowledge encompasses more than paradigmatic relations. We see it as knowledge which does not rely on any external events, which means that it is pretty impermeable to context[3]. Table 2 shows that this type of knowledge is also present in the AHFD, and it also suggests an organization of this knowledge.

Table 2. Static knowledge found in AHFD, categorized + KPs

Know. Type	Semantic Relation	Example
Composition	Part-of	An arm **is a part of** the body.
	Piece-of	A block **is a piece of** wood.
	Area-of	A beach **is an area of** sand.
	Amount-of	A breath **is an amount of** air.
Member-Set	Group	An army **is a large group of** people.
	Member	A letter **is one of** the symbols of the alphabet.
Human/ Animals	Relationship	Your cousin **is the child of** your aunt or uncle.
	Child	A lamb **is a young** sheep.
	Home	A hive **is a** home for bees.
Comparison	Like	A brush **looks like** a small broom.
Description	Name	An address **is the name of** a place.
	Material	Glass is what windows **are made from**.
	Function	A pen **is a tool to** write.
Intrinsic attributes[4]	Color	Toasts and chocolate **are brown**.
	Smell	It (onion) has a **strong smell** and taste.
	Size	A camera is a **small** machine that makes pictures.
	Taste	Lemons have a **sour taste**.

This investigation into the AHFD relates to much previous work on Machine Readable Dictionaries (MRDs). This research question of "how to find interesting knowledge" has been intensively studied, during the years in which doing knowledge acquisition from MRDs was fashionable [3,4,5,6,7]. With corpora now flowing on our desk, such research is often relinquished to the attic. We thought it was a good time to dust it off… Instead of starting from a clean slate when investigating corpora, we can certainly be inspired (even if the context is different) by research done on MRDs.

---

[2] Copyright @1994 by Houghton Mifflin Company. Reproduced by permission from THE AMERICAN HERITAGE FIRST DICTIONARY.

[3] Permeability to context, even for paradigmatic relations, in our opinion, cannot be absolute. The AHFD defines "prize" as *Prizes can be cups, ribbons, money, or many other things.* The notion of hyponyms (subclasses) of prizes is not really definable. There might be "typical" prizes, but nothing prevents just about anything to be a prize, depending on context.

[4] In Sowa (1984) a distinction is made between (a) attributes, holding for a concept without involving other concepts and (b) relations, which links a concept to other concepts of a domain. Given that distinction, the list given here of "intrinsic attributes" would not be considered a semantic relation.

Many researchers have been looking for KPs in definitions, but have referred to them more as "defining formulae"[5].

In a domain model, objects must be put in dynamic relationship to each other. This includes temporal and causal information. If information about time can be seen as a linear progression of state changes, information about cause-effect introduces a tree-like progression, leading to an immense number of possible state changes, as each node of the tree represents a cause-effect rule that could be apply. The possibility rather than the certainty of applying a rule reveals the intimate link between causality and uncertainty. This interaction between temporality, causality and uncertainty is what we refer to as **dynamic knowledge**.

Although the word dynamic brings to mind actions and therefore verbs, there is still a relation to nouns by the fact that some of them are causal agents. For example, the AHFD's definitions of "dust" and "tornado" involve causality. As such, "Dust can make you sneeze" and "Tornadoes can knock down houses and pull trees out of the ground". These examples relate to causal agent and are part of objects (nouns), somehow given their potential. Looking at verbs, we can call "intrinsic causality" the fact that some verbs intrinsically express transformations, or results or goals. We therefore look into the AHFD to find KPs for intrinsic causality and present our findings in Table 3.

Table 3. Intrinsic causality found in AHFD, categorized + KPs

Semantic Relation	Example
Result	Ash is **what is** left **(pp)** after something burns.
	Smoke **is made by** things that burn.
Cause	To kill **is to cause to** die.
	To pour **is to make** liquid go from one place to another.
	The window broke **when** a baseball went through it.
Transformation	To die means **to become** dead.

Table 4. Intrinsic temporality found in AHFD, categorized + KPs

Semantic Rel.	Example
Non-ordered parts	Exercise is running and jumping and moving your body around. **(list of –ing verbs)**
	The ground shakes and sometimes buildings fall **during** an earthquake.
Time-spanned event	A chase **is when** someone follows something quickly.
	A trip is **a time when** you travel somewhere.
Process	To roll is **to keep** turning over and over.
	To twist is to turn **around and around**.
Sequence	To dip means to put something in liquid **and then** to take it out quickly.

[5] Meyer [1] mentions that KPs have been called formulae, diagnostic frames or test frames, frames, definitional metalanguage and defining expositives, and knowledge probes. We refer the reader to Meyer [1] for all appropriate references.

The second type of dynamic knowledge mentioned above is temporal knowledge. Similarly to the idea of intrinsic causality, we can talk of intrinsic temporality. Some verbs embody within their meaning a succession of events, or minimally the notion that something happening spans over a period of time. We look at KPs to allow us for the discovery of such verbs in Table 4.

We have presented static and dynamic knowledge. A third type of knowledge will be present in a corpus, that of event-related knowledge. The usual who, where, what, when, how, why questions are part of events. An investigation into the AHFD will show that yes, many verbs do carry within their meanings some answers to these questions. Continuing with the intrinsic/extrinsic difference made above, we would say that these verbs contain intrinsic event knowledge. This is different from the long tradition of case role investigation [8], looking at extrinsic event knowledge. Some of these verb definitions are in Table 5.

Table 5. Intrinsic event-related knowledge

Semantic Relation	Example
Instrument	To bite means to cut **with** your teeth.
Method	To blow means to make a sound **by** pushing air.
Manner	To giggle is to laugh **in a** silly **way**.
Direction	To bow means to bend the body **forward**.
Path	To eat means to take food into the body **through** the mouth.
During	To dream means to imagine stories **while** you sleep.
Frequency	To practice is to do something **many times** so that…
Reason	Many people hug each other **to show that** they are glad.
Goal	To chase means to run after something **to try to** catch it.

Furthermore, some nouns are defined as place, agent or point in time. Table 6 shows a few examples. The noun definition therefore leads to a typical action and becomes an answer to who, where or when for that action.

Table 6. Intrinsic event-related knowledge expressed by nouns

Semantic Relation	Example
Agent	A barber **is a person who** gives haircut.
Location	An airport **is a place where** airplanes take off and land.
Point-in-Time	Birth is **the moment when** a person is born.

In conclusion, we have found in the definitions of a dictionary (which has traditionally been the knowledge repository for humans) three types of knowledge: (a) static knowledge of which the important fragment of paradigmatic relations is well known and well used in knowledge bases (b) dynamic knowledge comprising causal and temporal knowledge, and (c) event knowledge. Much research has been done to study many different aspects of knowledge, and it would be impossible here to refer to it all. Our modest contribution is from a definition analysis point of view, and was aimed simply at showing what is possible to find in dictionary definitions. The organization we suggest helps understand the type of information in the definitions, and will also help understanding the type of information often seek after in corpora.

3 Moving from dictionary to corpora

Now that we have looked in dictionary definitions for inspiration on types of knowledge to be discovered from text corpora, let us move on to discuss the actual activity of discovering Knowledge Rich Contexts (KRCs) in text corpora via the identification of Knowledge Patterns (KPs). To focus our discussion, we will look at one particular corpus on the subject of scuba-diving. The 1M word corpus covers different aspects of scuba-diving, such as medical, trips, equipment, safety, dangers, diving procedures, diving types, etc.[6]

Going back to the mention in the introduction of two main communities having an interest in KRC discovery, that is the Information Extraction (IE) and the Computational Terminology (CT) communities, we can highlight here a first distinction. Within the CT community, such a corpus on scuba-diving is of much interest as it is purposely assembled to contain informative texts (as opposed to narrative [9]). From those texts, terminographers will need to generate definitions for the important terms found, such as cave diving, overhead environment diving, decompression trauma, nitrogen narcosis, and barotrauma. Computational terminologists are interested in terms (concepts) [10,11] and semantic relations that will link these terms and allow them to generate good definitions [12,13,14,15].

In IE, the type of text is usually not specified, but the types of queries and required template filling have typically been of the type "who did what to whom, when, where, how and why", which presupposes narrative types of texts. For example, a text relating a diving accident occurrence could be transposed in a template form. Different systems have been built over the years for such purposes, such as AutoSlog [16], CRYSTAL [17], LIEP [18], PALKA [19], and ESSENCE [20], to name a few.

Relating now to the types of knowledge found in dictionary definitions, Terminological Knowledge Bases (TKB) tend to embody static knowledge, contrarily to IE being more interested in event-related knowledge[7]. The two meet with the dynamic knowledge, more specifically the causal knowledge via the causal agents, of interest in CT, and the how and why questions to be answered in IE.

Static knowledge is a good start for a TKB, and might be sufficient for terminographers to write definitions, but is not sufficient for the TKB to become a domain model. Some researchers [21,22] have started to investigate how KPs could help the discovery of causality, adventuring outside the static type to the dynamic type. Although not expected to be present in a MRD, as causality emerges from the interaction of the different agents within a system, we have argued elsewhere [23] that in fact causality takes root in terminology when we look at causal agents and the function relation. As for events, they are transients and are often not of much interest to store in a knowledge base, unless, these events embody some notion of generality. If they are not "one-time event", or events pertaining to specific characters, if they express how things "normally" are, they do have their place in a domain model.

---

[6] The author would like to thank Elizabeth Marshman for generating the scuba-diving corpus and to Ingrid Meyer for giving us access to it.

[7] Event-related knowledge in its extrinsic form is meant here as opposed to the intrinsic form seen in the definitions of the AHFD in section 2.

This illustrates again a main difference between CT and IE. In IE, there is an interest for one-time events, but in CT, there is an interest for generalization. The type of texts will be different, the KRCs of interest will be different, and the KPs will be different. Now, looking at these KPs, let us tackle the issue of generality. This comes as two questions: (1) is a given relation expressed similarly in dictionary and corpora? (2) do corpora domains influence the type of relation to look for beyond the ones identified in the MRD?

4.1 Expression of semantic relations – Issues of generality

Let us look at static and dynamic knowledge, and focus on the first question, the varied expression of a semantic relation. Only one or two KPs have been given for each semantic relation in the Tables of section 2, but the variety is much larger even within the AHFD. In the tradition initiated by Hearst [24], bootstrapping is used as a good way of finding alternate KPs for a relation given known concepts pairs connected through that relation. Our investigation on the scuba-diving corpus has not shown any unusual KPs for the relations of hypernymy and meronymy.

Table 7 shows a small sample of semantic relations with the KRC found including a KP. But a more extensive search, with multiple corpora would be required to do any fair assessment. Meyer [1] had mentioned variability across domains, but most often we are simply concern with variability whether it is within a domain or across. A very interesting work by [25] looks into the notion of variability of dependency structures instead of strings (making it more flexible) and relates this work to work on paraphrasing.

Table 7. Domain-independent information found in scuba-diving corpus

Semantic Relation	Example
Part-of	buoyancy-control, body positioning and propulsion techniques **are part of** both Cavern and Cave Diver training.
Hyperonymy	an air embolism **is another kind of** decompression illness
Cause	a lung over-expansion injury **caused by** holding your breath while you ascend.
Definition	The opening of the eustachian tube **is called** the ostium.
Function	Diazepam **is used to** prevent and treat oxygen convulsions and to control vestibular symptoms.

Now, moving on to the second question given above, Table 8 showing how the individual domain of scuba-diving contains its own domain specific relations forces us to answer yes. In fact, we have argued elsewhere [26] that any list used should depend on the application and that difference in many lists suggested are often issues of granularities, therefore suggesting a hierarchical organizational view of relations.

Table 8. Domain-specific information found in scuba-diving corpus

Semantic Relation	Example
Emergency measure	Pure oxygen **is first aid for** any suspected decompression illness
Symptom	The most common barotrauma **symptom** a diver experiences may be mild discomfort to intense pain in the sinus or middle ear.
Risk prevention	Keeping your SPG and high-pressure hose clipped to your left-hand side significantly reduces **the risk of** gauge damage entanglement.

Although outside the scope of this paper, it would be quite interesting to do an across-domain study to see if resemblance and differences can be merged within a hierarchical schema. Such integration attempts have been performed on concepts, such as in the work of [27] merging terms with the Wordnet [28] hierarchy. But to our knowledge such an across-domain hierarchical integration of semantic relations has never been performed.

4. Knowledge patterns: definition and productivity

If KPs provide our entry point into KRCs of the corpora, whether it is static, dynamic or event knowledge, we must have a flexible way of defining them. We will first present our view of defining KPs as including lexical, syntactic and semantic information. This expands on the more common lexical-syntactic view by including a semantic component. We then look at pattern productivity and discuss their use as entry points in the text.

3.1 Toward lexical-syntactic-semantic patterns

Lexical patterns found through string matching are the simplest form of KPs. They are useful but quite limited in their capacity to retrieve KRCs. Table 9 shows some examples of different types of knowledge (defined in Section 1) found with lexical patterns.

Table 9. Lexical information in KPs.

KP	Semantic Relation	Knowledge Type
is a kind of	Hyperonymy	Static – Paradigmatic
is a tool to	Function	Static – Usage
is to cause to	Causal	Dynamic-Causal
is a person who	Agent	Event-related
to show that	Reason	Event-related

To introduce flexibility, we move toward syntactic information. This simply assumes a link to a part-of-speech (POS) dictionary and does not necessarily require any

complex grammatical analysis. It does not even assume any kind of POS tagging which usually implies a disambiguation step. In our research, POS are used as tokens for pattern definition. The syntactic tokens add a lot of flexibility to the search without the burden of doing actual parsing. Table 10 shows some examples.

Table 10. Syntactic information in KPs.

KP	POS	Relation	Expected variations
is a *|a group of	Adjective	Group	large, small, eclectic
is a tool *|p	Preposition	Function	to, for
is to *|r make	Adverb	Causal	really, largely, principally
is what is *|v	Verb	Result	done, left, gained
is a *|n who	Noun	Agent	person, animal, doctor, student

Some POS provide a much larger search space than others. Nouns, for example which account for much more than fifty percent of the words in a dictionary, will generate a large set of answers. It is therefore important to make sure that such a POS is integrated in a pattern which restricts the set of possible nouns. Last row of Table 10 for example shows a pattern with the relative pronoun “who” which will exclude all objects from the preceding noun. On the other hand, a category such as preposition, since a closed-set category (function words), will be much more limiting.

The same way as the relative pronoun “who” above can be seen as restricting the semantic class of the previous noun, we can imagine that knowing the semantic class of a word can help restrict the possible patterns in which they can occur. The possibility of including semantic information in the KP assumes the availability of some lexical-semantic resource. Let us focus on two paradigmatic relations: hyperonymy and synonymy. The same as the syntactic token "noun" (*|n) could be replaced by any noun in the dictionary, a semantic token &home/^home could be replaced by any synonym/hyponym of home. Such semantic information can be found in a resource such as Wordnet [28]. One of main relation in Wordnet is the synonym relation (through the organization of words in synsets) and the hyperonym relation. Wordnet contains a lot of information about hyperonyms in the general common language. Table 11 shows some examples.

Table 11. Semantic information in KPs.

KP	Semantic Link	Semantic category	Expected variations
is a &home for	Synonymy	Home	house, living-place, roof
are ^colour	Hypernym	Colour	brown, blue, green
is a ^time when	Hypernym	Time	period, moment
is an &amount of	Synonymy	Amount	Quantity

Although each kind of token has a value when used alone Table 12 shows some possible interactions, and this is that type of search that would include lexical, syntactic and semantic information, providing a quite flexible way of getting access to KRCs.

Table 12. Interaction of lexical, syntactic and semantic information in KPs.

Relation	Expected variations
is *ld &amount *lp	is an amount for, is a quantity of, is a number of
*ln are parts *lp *la ^tree	branches are parts of a tree, needles are parts of a pine tree
*lv with ^instrument	write with pencil, draw with crayon
*ln is a *la ^animal	elephant is a large animal, canary is a small bird

4.2 Pattern Productivity

Although we can define KPs in a flexible way, we still need to investigate how much these KPs actually are the good guides or good entry points into KRCs. This leads us to the question of evaluation. The problem with evaluation is that it usually implies a task for which anticipated results exist, and against which an automated procedure can be compared. In an IE task, texts can be read manually and questions prepared in relation to the information in the text. For CT, it's not as clear what a good evaluation measure is, since the goal is not really to answer to singled out questions, but to "all" eventual questions by providing definitions for terms. So it is trying to get as much knowledge as possible. Still, researchers in CT wish to evaluate the productivity KPs and sometimes use a measure of noise and silence, the complements of precision and recall traditionally use in IE. The calculations are prone to errors since they require a human manually looking through a corpus of texts, and for a singled out semantic relation, for a singled out KP, evaluate noise and silence. This process, even if informative, is also questionable, as it is totally depended on the corpus used, and therefore makes the measures very subjective. Without going manually through the whole corpus, we can minimally evaluate how noisy a pattern is. We can also evaluate what we will define as *relative productivity* by comparing the number of good KRCs identified by all the patterns used and assuming a uniform distribution across patterns.

Focusing on the function relation, Table 13 presents a list of patterns with their frequency of occurrence, noise and relative productivity. We can note in the second to last pattern the use of a wildcard "design*" to cover: design, designs, designed. The last pattern uses POS tag of preposition "used *lp" allowing the retrieval of 47 patterns distributed as the following coverage: used to (15), used in (9), used by (8), used for (5), used with (4), used as (2), used up (1), used like (1), used on (1), used after (1).

A KP with a relative productivity as good as all other patterns would give 100%. Let us call NbExp the number of expected occurrences given a uniform distribution across patterns, the relative productivity would be calculated as: Number of good occurrences – NbExp / NbExp.

A pattern is noisy if sometimes it is indicative of a semantic relation and sometimes not. A different form of ambiguity is when a single pattern can lead to different possible semantic relations. This is the problem we face when we investigate prepositions as patterns. In Table 14, we show a few prepositions and their ambiguity. We

went back to the AHFD for the examples, as they provide very clear and simple examples.

Table 13. Statistics on different KPs for the function relation. (i – Total number of occurrences of pattern, ii – Number of occurrences NOT showing a KRC, iii -- Percentage of occurrences NOT indicating "function", iv – relative productivity)

Pattern	(i)	(ii)	(iii)	(iv)	Positive Example	Negative Example
serve to	1	0	0%	-91%	Reef hooks (…) may **serve to** limit damage to both reef and diver.	
Useful for	1	0	0%	-91%	Some chemotherapy is **useful for** marine animal injuries.	
made to	2	0	0%	-82%	small incision needs to be **made to** extract the spine	
Intended for	1	0	0%	-91%	regulator second stage **intended for** use as an octopus.	
Design* to	28	2	7%	136%	our lungs are **designed to** breathe gas	They are similar in **design to** the active-addition oxygen re-breathers
used *|p	47	11	23%	227%	drugs like those **used to** control cold symptoms	I **used to** go so far as to tell people ….
Total	80	13	16%			

Table 14. Highly ambiguous KRCs

KRCs	Semantic Relation	Example
for	Function	A carpenter has a box **for** his tools.
	Recipient	I bought this book **for** you.
	Direction	People can reach **for** the sky, but they can't touch it.
	Duration	We played baseball **for** 2 hours.
in	Location	Fish swim **in** the water.
	Point-in-time	Ted's birthday is **in** August.
	Manner	Ants live **in** large groups.
	Containment	The people **in** this story are Lisa and David.
of	Part-of	Branches grow out from the branch **of** a tree.
	Containment	Paul was carrying a pail **of** water.
	Point-in-time	The time is 10 minutes **of** four.
	About	Steve drew a picture **of** his brother.
	Possession	You can see the work **of** many artist in a museum.
	Material	We made a border **of** stone around the garden.
with	Containment	Joe's hamburger can **with** onions on it.
	Part-of	A giraffe is an animal **with** a long neck.
	Instrument	Brian dug a hole **with** a shovel.
	Manner	Attention is looking and listening **with** care.
	Accompaniment	To march **with** someone means to take the same size steps at the same time.

This case is interesting as ambiguous prepositions might be very problematic in a IE system, and unfortunately prepositions are some of the most important indicators of event-knowledge, but it might not be so problematic in a CT system, since the terminographer is looking for any interesting information, and is usually restricting the contexts of search to ones containing a term of interest.

It is important though in an interactive system to reduce the load on the user as much as possible [29]. Some even suggest that the user should never be bother with defining patterns KPs but should just be given KRCs and say yes-no (implying that a pattern learner listens in the background) [30]. Depending on how well the learner performs, this might be more or less burden on the user, if he has to constantly be asked yes/no on inappropriate KRCs.

5 SeRT – Knowledge Acquisition Tool

In this section, we briefly describe SeRT (Semantic Relations in Text). SeRT is a Knowledge Acquisition tool which relies at its core on the hypothesis that explicit knowledge is acquired via the discovery of semantic relations, and these semantic relations are expressed via KPs.

In [31] further details are given about some of SeRT functionalities included in an early version, but as an overview, we can say that SeRT has (1) a Term Extraction module (finding a list of terms in a corpus), (2) a Semantic Relation Search module, which implements in a limited way the lexical-syntactic-semantics search (3) a storage capacity comprising two aspects a) a list of semantic relations and their patterns, b) a list of semantic relations and the concepts they link to generate a knowledge base of the extracted information, (4) a visualization module for the knowledge base. For future work, we will focus on the integration of the Wordnet resource with our search module to augment the semantic aspects. For now, the only semantic relations between concepts known are the ones that are iteratively added to the knowledge base.

SeRT is being developed as an aid to a computational terminologist or a knowledge engineer. In a highly interactive way, it gives the human user a flexible tool to discover Knowledge Rich Contexts (KRCs) through the definition of Knowledge Patterns (KPs). A KRC can be looked at within a small context (window with 4 words on each side) to quickly eliminate invalid sentences. A chosen KRC can then be seen in a large context (two sentences) to be looked at by the human user for identification of concepts and semantic relations to be entered in the knowledge base.

To help a new user get started, a list of semantic relation is available to the user, and for each, a list of knowledge patterns is also available. As the user goes through the knowledge acquisition process, the new patterns and new semantic relations identified are added.

The use of SeRT allows the construction of a Knowledge Base on a specific domain, such as the extracts presented in Section 3, on the topic of Scubadiving.

6 Conclusion

Our purpose in this research was to investigate Knowledge Patterns (KPs), as they are the key entry points to the discovery of Knowledge Rich Contexts (KRCs), contexts containing valuable information. We discuss the notion of information value by looking at the needs of two communities interested in knowledge acquisition, the Information Extraction and the Computational Terminology communities. A small children's dictionary served as the starting point to make the link between KPs and semantic relations to be included in a knowledge base. This also allowed us to regroup the type of knowledge present in the definitions in three categories: static, dynamic and event-related, and to make an attempt at characterizing which type is of more value for each community. We further looked at text corpora where further domain-specific relations will need to be defined as well. We provided some ideas on the definition and the productivity evaluation of the KPs. The definitional framework we suggest combines lexical, syntactic and semantic information. The evaluation framework we suggest limits the human burden by considering only the KRCs instead of all the sentences in a corpus, and gives a relative notion of productivity. Finally we have presented SeRT which is an interactive tool for knowledge acquisition. Further investigations, as mentioned before will include integration of the semantic capabilities of SeRT with Wordnet, as well as further refinement to pattern definitions, to include even more complex patterns, such as (as NP1 vb1, NP2 vb2) && (vb1 antonym vb2) [32] which includes logical operators. This will render SeRT an even more flexible and powerful tool for helping humans in their process of knowledge acquisition and organization.

References

1. Riloff, E.: An empirical study of automated dictionary construction for information extraction in three domains. In: Artificial Intelligence 85, (1996) 101-134
2. Meyer, I.: Extracting Knowledge-rich Contexts for Terminography: A Conceptual and Methodological Framework. In: Bourigault, D., Jacquemin, C., and L'Homme M.C. (eds): Recent Advances in Computational Terminology, John Benjamins, (2001) 279-302
3. Chodorow, M.S., Byrd R.J., and Heidorn, G.: Extracting semantic hierarchies from a large on-line dictionary. In: 23rd Annual Meeting of the Association for Computational Linguistics, (1985) 299-304
4. Ahlswede, T., and Evens, M: Generating a relational lexicon from a machine-readable dictionary. International Journal of Lexicography 1(3), (1988) 214-237
5. Dolan, W., Vanderwende L, Richardson, S.D.: Automatically deriving structured knowledge bases from on-line dictionaries. In: The First Conference of the Pacific Association for Computational Linguistics, Vancouver, (1993) 5-14
6. Wilks, Y., Fass, D., Guo, C.-M., McDonald, J., Plate, T., and Slator, B.: Providing machine tractable dictionary tools. In: Pustejovsky, J. (ed.): Semantics and the lexicon, chapter 16, (1993) 341-401
7. Barrière, C., and Popowich, F.: Building a Noun Taxonomy from a Children's Dictionary. In: Gellarstam M. et al. (eds): Proceedings of Euralex'96, Gotebord, (1996) 27-34

8. Fillmore, C.: The Case for Case, In: Bach, E., and Harms, R. (eds): Universals in Linguistic Theory, New York, Holt, Rinehart and Wilson (1968)
9. Kintsch, W. and van Dijk, T.A.: Toward a model of text comprehension and production, Psychological Review, 85(5), (1978) 363-394
10. Kageura, K., and Umino, B.: Methods of Automatic term Recognition: A Review. Terminology 3(2), (1996) 259-290
11. Cabré Castellvi, M.T., Bagot, R.E. and Palatresi, J.V.: Automatic term detection: A review of current systems. In: Bourigault, D., Jacquemin, C., and L'Homme M.C. (eds): Recent Advances in Computational Terminology, John Benjamins, (2001) 53-87
12. Biebow, B. and Szulman, S.: TERMINAE: A linguistics-based tool for the building of a domain ontology. In: 11th European Workshop on Knowledge Acquisition, Modeling and Management (EKAW'99), Dagstuhl Castle, Germany, (1999) 49-66
13. Bowden, P.R., Halstead, P., and Rose T.G.: Extracting Conceptual Knowledge from Text Using Explicit Relation markers. In: Shadbolt, N, O'Hara K., and Schreiber G. (eds): Proceedings of the 9th European Knowledge Acquisition Workshop, EKAW'96, Nottingham, United Kingdom, (1996) 147-162
14. Condamines, A. and Rebeyrolle, J.: CTKB: A Corpus-based Approach to a Terminological Knowledge Base. In: Bourigault, D., Jacquemin, C., and L'Homme, M-C. (eds): Computerm'98, (1998) 29-35
15. Meyer, I., Mackintosh, K., Barrière, C., and Morgan T.: Conceptual sampling for terminographical corpus analysis. In: Terminology and Knowledge Engineering, TKE'99, Innsbruck, Austria, (1999) 256-267
16. Riloff, E.: Automatically constructing a dictionary for information extraction tasks. In: Proceedings of the Eleventh National Conference on Artificial Intelligence, (1993) 811-816
17. Soderland, S., Fisher, D., Aseltine, J., Lehnert, W.: CRYSTAL: Inducing a conceptual dictionary. In: Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (1995) 1314-1319
18. Huffman, S.B.: Learning information extraction patterns from examples. In: Lecture Notes in Computer Science, Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing, (1996) 246-260
19. Kim, J., and Moldovan, D.: Acquisition of linguistic patterns for knowledge-based information extraction. In: IEEE Transactions on Knowledge and Data Engineering, 7(5), (1995) 713-724
20. Català, N., Castell, N., and Martin, M.: Essence: A portable methodology for acquiring information extraction patterns. In: ECAI'2000, Proceedings of 14th European Conference on Artificial Intelligence, (2000) 411-415
21. Garcia, D. : Structuration du lexique de la causalité et réalisation d'un outil d'aide au repérage de l'action dans les textes, Actes des deuxièemes rencontres - Terminologie et Intelligence Artificielle, TIA'97, (1997) 7-26
22. Barrière, C.: Investigating the Causal Relation in Informative Texts, Terminology 7(2) (2001) 135-154
23. Barrière, C., and Hermet, M.: Causality taking root in Terminology. In: Proceedings of Terminology and Knowledge Engineering, TKE'2002, (2002)
24. Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora. In: Actes de Coling'92, Nantes. (1992) 539-545
25. Lin, D. and Patel, P.: Discovery of Inference Rules for Question Answering. In: Natural Language Engineering, 7(4). (2001) 343-360
26. Barrière, C.: Hierarchical Refinement and Representation of the Causal Relation, Terminology 8(1), (2002) 91-111

27. Alfonseca, E., and Manandhar, S.: Extending a Lexical Ontology by a Combination of Distributional Semantics Signatures. In: EKAW 2002, LNAI 2473, Springer-Verlag. (2002) 1-7
28. Miller, G.A.: WordNet: a lexical database for English, Communications of the ACM, 38(11), (1995) 39-41
29. Gil, Y. and Ratnakar, V.: IKRAFT: Interactive Knowledge Representation and Acquisition from Text. In: EKAW 2002, LNAI 2473, Springer-Verlag. (2002) 27-36
30. Brewster, C., Ciravegna, F., Wilks, Y.: User-Centred Ontology Learning for Knowledge Management, NLDB 2002, LNCS 2553, Springer-Verlag Berlin Heidelberg (2002) 203-207
31. Barrière, C., and Copeck T.: Building a domain model from specialized texts. In: Proceedings of Terminologie et Intelligence Artificielle, TIA'2001, Nancy, France (2001) 109-118
32. Moldovan, D.I. and Gîrju, R.C.: An interactive tool for the rapid development of knowledge bases. In: International Journal on Artificial Intelligence Tools, 10(1,2) World Scientific Publishing Company (2001) 65-86