Variation and Semantic Relation Interpretation: Linguistic and Processing Issues

Nathalie Aussenac-Gilles, Anne Condamines

Variation and Semantic Relation Interpretation: Linguistic and Processing Issues

Nathalie Aussenac-Gilles[1], Anne Condamines[2]

[1]IRIT, Université de Toulouse, CNRS, France
aussenac@irit.fr
[2]CLLE-ERSS, Université de Toulouse, CNRS, France
anne.condamines@univ-tlse2.fr

Abstract. Studies in linguistics define lexico-syntactic patterns to characterize the linguistic utterances that can be interpreted with semantic relations. Because patterns are assumed to reflect linguistic regularities that have a stable interpretation, several software implement such patterns to extract semantic relations from text. Nevertheless, a thorough analysis of pattern occurrences in various corpora proved that variation may affect their interpretation. In this paper, we report the linguistic variations that impact relation interpretation in language, and may lead to errors in relation extraction systems. We analyze several features of state-of-the-art pattern-based relation extraction tools, mostly how patterns are represented and matched with text, and discuss their role in the tool ability to manage variation.

Key-words. Pattern-based relation extraction, lexico-syntactic patterns, semantic relation, tool comparison, pattern variation

1 Introduction

Knowledge representation in most onto-terminological resources (ontology, terminologies, thesaurus…) refers to relational networks and more precisely, concepts related by binary relationships, either hierarchical like hypernymy, or not, like causality, possession, qualities or properties. When building this representation from text, the state-of-the-art talks about 'concept extraction' and 'relation extraction', which assumes that language produces direct evidences of these knowledge structures. Many relation extraction tools parse texts as if binary relationships could directly be associated with some predefined sequences of syntactic or lexical tokens. Among all possible implementations, we focus on pattern-based approaches. When using patterns, there is a strong temptation to look for schemes (i.e. noun_phrase/verb/noun_phrase sequences) in text and to map them to concept/relation/concept triples. Nevertheless human interpretation relies on much more subtle and complex linguistic associations and on hidden background knowledge. So far, a large variety of concept and relation extraction tools have been implemented for research

Aguado de Cea et al. (Eds.): Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012), pp.106-122. 19-22 June 2012, Madrid, Spain

purposes. They may either assist concept and relation identification in text, or run the entire process from language analysis to knowledge modelling.

It is not trivial to identify textual contexts that may reveal semantic relations, and then to decide of the relations that can be represented. Accounting for the linguistic features that contribute the relation interpretation is already complex, regardless of any automation perspective.

In this paper, we browse the various linguistic and technical issues that may arise at the various stages of this process. We discuss how existing studies or systems carry out these stages and manage (or not) to overcome these issues. Examples of such issues are: what are the relevant linguistic marks to identify concept/relation/concept triples? What are the observable variations that may affect the interpretation of semantic relations? How far can tools assist this process? How could tools better anticipate these variations? In first two sections we adopt a linguistic point of view. The next two sections aim to connect the issues raised by the linguistic observations with some of the features of pattern-based relation extraction systems. We conclude with perspectives to better support relation extraction with software tools.

2 Linguistic variation in lexical relation interpretation

Knowledge rich contexts are parts of text which may be interpreted by human as giving information about a term and even by giving it under structures such as term/relation/term. But this interpretation process faces two types of problems which make it hard to carry out by a software tool.

2.1 Variation affecting relation interpretation

In her paper [1], Meyer speaks about *knowledge rich contexts (KRC)*. The KRC term notifies that some parts of texts are richer than others concerning knowledge but it does not specify that these parts may be represented under a term/relation/term triple. The linguistic characterisation of KRC is complex [2], [3], [4]. It requires analyses that go beyond the sentence level to take paragraphs into account; it requires not only to justify the relation label but also to locate the related terms. A valid interpretation and triple definition would require detailed and deep linguistic analyses. More generally, sentences that contain standard patterns are not very frequent or they may lead to some semantic relations that are not relevant to be included in a knowledge model.

2.2 Instability within discourse

Spotting term/relation/term in texts is not easy. The role of language, even in specialized fields doesn’t consist in referring to knowledge by using explicit

triples. In addition, speakers do not always intend to express domain knowledge. Some definition contexts are very implicit and may be considered as such only by the linguist, the terminologist, or the knowledge engineer who wants to define concepts. We identified the following four categories of problems, that is to say differences between an *a priori* interpretation of discourse functioning and the actual term-relation-term structure.

One of the terms-concepts is missing. In case of anaphora, a hypernymy relationship may link a first term in a phrase and another one in another phrase. For example in *… a car. This vehicle…*, *this vehicle* refers to the same object as *car* in the previous sentence and there is a hypernymy between *vehicle* and *car*. But in some cases, it is not possible to identify a term with which the second term may be related such as in:

(1) The configuration units to be modified must be identified […]. The result of this activity is the drawing up of the modifications file.

Because *this activity* refers to something presented in the previous sentence, *Activity* can be identified as a hypernym. But no term in the previous sentence may be understood as a hyponym: the hyponym is meant by the whole phrase "the configuration units to be modified must be identified".

T1 and T2 have not the same grammatical nature. In ontologies, most of the concept labels are noun phrases, but verbs or adjectives can also be found. Depending on the relation nature and on the term meaning, it may be odd or even erroneous to connect concepts with labels of a different grammatical nature (a noun and a verb for example). Though, such a connexion may exist in discourse. In example (2)

(2) The numbering of cables consists of identify and number each cable for an electrical cupboard.

the first term (*numbering of cables*) is a noun phrase while the second one (*identify and number…*) is a verb phrase. The pattern *consists of* is well known as playing a role in definition so this context may be identified as expressing an equivalence between two terms. The first term probably is a kind of elliptical form of the second one: *numbering of cables* is equivalent to *identifying and numbering each cable for an electrical cupboard*.

Pattern and T2 are present in the same word. In some cases, a word matches both the pattern under prefix form and term2 of the triple. From example (3) one may deduce that *calcium* generally is a constituent of *bone*.

(3) This bone is decalcificated.

The privative prefix *de* marks mainly that a component, generally present in the whole object, is suppressed (such as in *decaffeinated*). So, this prefix may be used as a meronymy mark. But it is not always the case. For example, from

depolarization, it is not possible to deduce that there is a meronymic relation between *polarization* and the object which is depolarized. Morphological features are not easy to describe and even more to introduce in a tool.

For examples (1), (2), and (3), discourse analysis is required to build a term/relation/term triple with the help of human interpretation. So far, tools are not suitable to account for such discourse behaviors. Some linguistic phenomena have not been really described from a linguistic perspective and even less with the aim of building ontologies or terminologies. There is a problematic lack of linguistic studies usable for designing tools.

2.3 Variation when building a representation

Six phenomena are examined in this paragraph: sub-relationships, polysemy of patterns, multiple possible meanings, implicit relationship, rhetorical effect and indirect interpretation.

Polysemy of patterns. Some patterns may correspond to two different relationships not necessarily specific of another relationship [1]. It is the case of *comme* in French (corresponding to *as* in some English examples) [5]. In (4), this pattern reveals a hypernymy (decorative flower is hypernym of rose) wheres as in (5), the relation between the two nouns is rather co-hyponymy (Rose and orchid are both hyponyms of flower):

(4) La rose comme fleur de décoration est très appréciée (The rose, as a decorative flower, is very much appreciated).

(5) La rose comme l'orchidée sont très appréciées des clients. (Roses as well as orchids are very much appreciated by customers.)

In some cases, the difference between interpretations may be explained by syntactic context and by textual genre. But again, this characterization requires fine-grained and time-consuming analyses.

Multiple possible meanings: class vs instance. According to the aim, the same sentence can get different interpretations either at the class or the instance layer. For instance,

(6) Roses are Marie's favorite flowers

can be interpreted at a general level, to learn the <*rose*, sub_class_of, *flower*> triple, or < *person*, hasFavoriteFlowers, *Roses*>. The two interpretations sound correct but the first one (the hypernymic relation) will be preferred to be integrated in an ontology. But the same sentence could also provide a probe for a relation between instances like <*Marie*, hasFavoriteFlowers, *rose*> where *rose* could be an anonymous instance of the *Rose* class.

Implicit relationship. In some cases, a pattern that systematically means a relationship can have a better interpretation which refers to another relationship. That may be the case for example with the succession relation that can be interpreted as a causative one [4], [6]. Example (7) expresses a succession to be interpreted as a cause: rain stopping causes the beginning of the show:

(7) The show will begin when the rain will stop.

From a linguistic perspective, these cases are well known. Nevertheless, in some sentences *when* has to be associated only with a temporal interpretation (succession or concomitance) as in (8):

(8) It was very cold when I went to New York last year.

Rhetorical effect. As parts of speech, patterns may be involved in rhetorical processes. Let's examine the sentence below.

(9) The component integration phase may begin when all the software elements have been implemented.

This example should be understood as meaning that software elements must first have been implemented for the component integration phase to begin. Therefore the expressed relation looks like a temporal relation but it must be understood as a conditional one. So, what is presented as just temporal information has to be considered as an injunction. The aim of using a temporal connective is to weaken the order effect. In examples (7), (8) and (9), the same pattern *when* may be understood either as a temporal, a causative or a conditional one. Some linguistic elements of the context may be used in order to disambiguate these possible interpretations but in some cases, the whole situation (linguistic and extra-linguistic context) is necessary in order to obtain the good interpretation.

Indirect interpretation. This case is very interesting because it highlights the fact that linguistic and knowledge engineering needs may lead to different descriptions. This is the case with *chez* (that may be translated by *among* in English). Some sentences containing *chez/among* may be used in order to build a meronymic relation as in (10), where *it may be understood that there is a meronymic relation between* nose *and* colobines.

(10) Among the colobines, the nose juts out over the upper lip.

As described in [5] and [7], this interpretation appears in texts of didactic origin dealing with natural science: in such texts more than 50% of the sentences where *chez* occurs contain a meronymic relationship.

But it is not true to say that *chez* systematically leads to a meronymic relation: nobody spontaneously produces this preposition with a meronymic meaning and its etymology (from the Latin *casa* (house)) is in no way linked to such an interpretation. A detailed analysis shows that this preposition occurs in structures in topic position (at the beginning, the middle or the end of

the sentence), i.e. in structures that introduce a new referent into the discourse. In didactic natural science texts, what is often said about these new referents (animals or plants) has to do with their anatomy or composition. Thus *chez* may be used as a sort of clue instead of really a complete pattern.

Multiple binary relations. A final difficulty arises when looking for multiple binary relations or n-ary relations. Compared to what actually occurs in text, concept-relation-concept triples are not very convenient to represent n-ary relations. For example, it is complex to build the patterns that would identify a communication n-ary relation [8]: NP1 (*person) communicates NP2 (information) to NP3 (person) through NP4 (mediaM) as in*

(11) Each subdivision transmits to CIGT a form related to complete site.

In this example, the three verb arguments are intrinsically tied, it appears clearly that binary relations are restrictive and do not convey the fact that the 3 binary relations have to co-occur.

2.4 Textual Genre and Pattern Meaning

It is obvious that the interpretation of some conceptual patterns is genre-dependant [5]. In other words, the probability for a word or a structure to be interpreted as a conceptual pattern is not equivalent in every text. Here is just the example of *avec* (*with*) and the results obtained from texts belonging to different genres. The corpus gathers texts belonging to five genres:

- A Zola novel: *Germinal*, 210,000 words, noted GER.
- A scientific handbook: "*Manuel de géomorphologie*", (Geomorphology handbook) 206,700 words (abbreviated as "GEO").
- A toy catalogue (*Catalogue de jouets Leclerc*), 93,000 words (T.C).
- Real estate adverts collected from 3 web sites, 22,600 words (P.A).
- Itinerary descriptions (a corpus constituted for the purpose of a psycho-linguistic study), 48,000 words (ITI).

The table 1 below presents the results of the study.

	GER	GEO	T.C	P.A	ITI
Avec	667	432	236	185	114
Meronymic *avec*	3%	12.7%	68.2%	76,2%	64,6%

Table1: Quantitative Results for *avec*

Two groups of corpora can be identified. In the first one (GER and GEO) the number of meronymic *avec* occurrences is very low. In the second one (T.C., P.A., ITI), the number of meronymic *avec* occurrences is very high. This observation warrants the claim that *avec* may be considered as a conceptual pattern with a high probability within this second group of corpora.

3 Pattern-based Relation Extraction Software

3.1 Automating the Search for Semantic Relations

During the last 20 years, the automatic search for semantic relations has been the goal of studies in information extraction [9], terminology collection from corpora [10], [11], [12], and ontology engineering from text [13], [14], [15], [16], [17], [18]. Semantic relations may be either directly provided by domain experts, acquired from text or reused from existing lexical or semantic resources. We focus on systems that build semantic relations thanks to pattern matching because they are particularly efficient to identify domain-specific relations and to label them precisely. We evaluate their ability to capture the variations presented in section 2.

Searching for relations thanks to pattern occurrences in text relies on several foundational assumptions [19]: (i) relation expressions are regular enough to be anticipated; (ii) relations can be found with similar formulations in any corpus; (iii) relation expressions match sequences of lexical entries and grammatical categories; (iv) in a body of text, a linguistic pattern is interpreted with a unique meaning. These hypotheses are being refined to find out efficient implementations that reduce noise and improve recall.

One limitation of pattern-based relation extraction is its low productivity: searched texts may contain very few explicit formulations of relations; depending on the text genre, patterns may occur very scarcely. Handbooks and lecture notes are acknowledged as adequate document genres for hypernymy and definition relation extraction. But novels for instance contain very few occurrences of Hearts's hypernymy patterns [19]. Pattern matching has a moderate success because of variation, which lowers recall: relations occur with different meanings or formulations, so that patterns miss many occurrences. From now on, we will adopt a knowledge engineering perspective and consider patterns not only as means to account for linguistic phenomena, but also as tools to get linguistic clues of the knowledge to be represented.

3.2 Tools for pattern-based search for semantic relations

Our study relies on several state-of-the-art surveys [20], [21], [22], [23]. Each of these papers reports a dozen of different tools, among which we identified four types of software that assist relation mining in text.

Text analysis platforms enable the deployment of natural language processing applications by combining basic components that apply at different

linguistic levels. For instance, Gate[1] includes the A**NN**otations **I**n **C**ontext (ANNIC) plug-in to identify concepts and relations thanks to rules that implement patterns. ANNIC can be used to annotate a corpus or to query annotation contexts [24]. LinguaStream[2], Alvis[3] or UIMA[4] are other such platforms.

Independant relation extraction tools: tools like Prométhée [25], Caméléon [14] or RelExt [26] are dedicated to relation extraction either with or without the identification of related terms. [27] proposes a tool to extract taxonomies from text, Terminoweb [10], Expresso [17] and Snowball [9] are able to identify together a relation type and related terms.

Ontology engineering platforms that use text as knowledge sources: Platforms such as Text-To-Onto [12], OntoLearn [17] or Terminae [28] support a methodology and include tools to look for terms and semantic relations.

Specific relation extraction tools: tools may be dedicated to a particular type of application or domain such as bioinformatics, for instance PASTA [29], RelationAnnotator [30], works by [31], [32], or [33].

Learning based tools: these tools exploit knowledge models or lexical resources (like WordNet) that provide pairs of related terms to learn new patterns from very large corpora. Because hypernyms are more easily observed in corpora, many of these tools learn taxonomic relations like [34] or TaxoLearn [35]. Other relations like causality, parthood or domain specific relations are searched by [36], [37] or with Prométhée [25].

4 Variation and Pattern-based Relation Extraction

In the following, we will identify the help that a tool can provide to relation extraction. Automatic extraction is easier in systems dedicated to a restricted number of relation types or to a specific kind of corpus. The internal representation of patterns as well as the ability to adapt them impacts the system ability to manage variations in relation formulations, but it cannot be predicted from the degree of automaticity. Some systems enable pattern learning so that they can adapt to each new corpus or domain, in such a way that automation contributes to better anticipate variations and to define more efficient patterns.

[1] http://gate.ac.uk
[2] http://www.linguastream.org/
[3] http://www-lipn.univ-paris13.fr/~hamon/PlateformeTAL-ALVIS/index.html
[4] http://uima.apache.org/

4.1 Anticipating Variation during Pattern Definition or Learning

Pattern definitions can be adapted to variation in systems that allow pattern reuse, or the search for domain specific relations with ad-hoc patterns.

Pattern Definition Assumes Stability. There is a strong paradox in trying to account for variation during pattern definition. Patterns try to account for regularity of relation expressions with regular expressions. This hypothesis is even stronger for "generic" patterns, like Hearst's patterns [19]: hypernymy is supposed to be captured by the same set of patterns in any corpus whatever its domain and genre. Nevertheless, practical reuse of generic patterns shows that the hypothesis is reductive. Patterns may be more or less reliable and they may reveal relations with a different meaning from the one expected. Results of matching a set of patterns on several corpora prove to be very useful when reusing these patterns. Such lists can be found in Caméléon for French, in Terminoweb for English and French, in Skeleton [37] for Catalan.

A first way to take variability into account at this stage is to allow for generic pattern reuse, evaluation and adaptation to the corpus, like in Caméléon or Terminoweb. Flexibility is increased by identifying corpus-specific patterns. Then users should be assisted in their definition, evaluation and adaptation. When defining a pattern, human interpretation is required to abstract relevant features from corpus sentences, to qualify the meaning of the relation and to select and represent a relevant pattern. The system can provide a database of reusable patterns and an interface for pattern definition or modification. Statistics about pattern use (number of occurrences, precision and recall in other corpora) are useful to guide pattern evaluation, selection or reject.

Pattern Learning from Corpora. Supervised machine learning algorithms offer promising perspectives to exploit abstract regularities from tagged corpora. Similar algorithms can be run to learn either semantic relations and taxonomies from textual evidences like [38] or [39] do it, or lexico-syntactic patterns. A variety of techniques can be used to « induce recurrent patterns » [22]. The ASIUM [38] pioneer tool classifies the contexts of each verb and abstracts patterns from them. Many such approaches adapt the DIPRE algorithm by Bri [21] like Prométhée [25], WWW2rel [21], [40] or [41]. The learning process exploits pairs of related concepts and their labels in corpora. The learning cycle consists in (1) building patterns for each context where concept pairs occur in the same sentence, (2) generalizing similar patterns by abstracting each slot, (3) evaluating learned patterns to avoid over generalization. For instance, in [42] Wordnet is used to search for causal relation patterns in a web corpus. Patterns are expected to match the <NP1 verb NP2> structure where NP1 and NP2 belong to related synsets. For each sentence containing NP1 and NP2, a <NP1 verb NP2> pattern is defined after manual

validation to filter out noisy patterns. An alternative algorithm used in Snowball [12] infers patterns made of surface grammatical features (like POS) that generalize the largest number of relation instances. Expresso [17] uses a clustering algorithm to abstract related classes from the terms found in each relation occurrence. In any case, learning patterns requires not only pairs of related concepts as bootstrapping data, but also large corpora with many relation occurrences. For this reason, this method cannot be used to look for domain specific relations in small corpora of technical documents. Moreover, pattern learning assumes that learned patterns will have the same stable interpretation all over the corpus, and that all their occurrences will mean the same type of relation, which is often a too strong hypothesis.

Pattern Evaluation on Corpora. According to [34] a reliable pattern is one that matches a large set of documents with a high precision, even though its recall is low. Evaluating a pattern reliability requires to estimate the quality of the concept pairs found in a corpus thanks to this pattern. Measures like mutual information between the corresponding terms, Kappa measure [21], or those listed in [15], can be used to check each learned relations. Related terms should have a stronger correlation value. Pattern quality is estimated by combining kappa values of each sentence extracted with this pattern. Reliability tends to promote precise patterns and leads to define numerous variants to account for linguistic variations. But pattern quality and reliability are not intrinsic pattern features [14]: they depend on the corpus domain and genre. An experiment on 8 corpora and 30 patterns showed that the same pattern can be frequent and precise in one corpus, and very noisy or rare in other ones.

4.2 Pattern structure and its relation with variation management

Pattern is a generic word that accounts for various kinds of structures with a variety of implementations. Each type of structure assumes a particular stability and ability to handle variation, and makes pattern definition and matching more or less complex. Here are some of the mostly used structures:

***word1* [+ *word2*].** Patterns made of lexical forms only focus on the informative part of a pattern, i.e. *chez* (in (15)) can be a pattern for meronymic relations. Verbal patterns in [39] are of this kind. Although easy to build up from lexical entries, they are not able to find the related terms and they lack abstraction: a large set of patterns is required to account for little variations.

A term1 B term2 C, where **term1** and **term2** are the related terms and A, B, and C characterize their lexical or grammatical context. In Caméléon [14], Marshman's work [4] and TerminoWeb [10], terms and context items can be either POS, lexical or semantic classes, empty words, information about their

localization, etc. Simplifications omit term1 and term2, or A and C. In Expresso [17], patterns conform to a similar shape: **ENTRY**/NP is/VBZ a/DT type/NN of/IN **TARGET** where ENTRY is term1 and TARGET is term2, NP is a noun phrase, DT a determiner, VBZ a verb and IN a preposition.

In these first two cases, all the items that form a pattern are of equal importance. They are searched in linear order. Anticipating the variation presented in section 2 requires to list all possible formulations and to define as many patterns as needed. An alternative solution gives more weight to the most significant parts of the pattern, key features that form its core, and less weight to secondary items. Stability is more likely to apply to the core, that is expected in every occurrence, whereas variation affects the secondary items, that are likely to be optional without affecting the relation meaning. Setting up secondary items has an operational influence on the pattern efficiency but it does not change its semantics. The next three structures illustrate this option.

***Left* term1 *Middle* term2 *Right,* where** *middle* is considered as more important than *Left* and *Right.* Here, the system will first look for the searched terms or their semantic categories and then, it matches *Middle* to their context. In SnowBall [9], patterns have the following shape:

(<**left**>, <LOCATION, Seattle>, <**middle**>, <ORGANIZATION, Boeing>, <**right**>)

Today's merger positions Seattle **–based** Boeing **the largest aircraft manufacture**

A similar structure is used for pattern definitions in [39]: <Left Verb_Def Nexus Right>, where the verb (Verb_Def) and its potential modifier (Nexus) are considered as more important than the terms in relationship which appear inside right and left fragments.

IF Initial clues THEN R (IF Contextual clues THEN R(Arg1,Arg2)) : Implemented in systems like Coatis [6] or ContextO [44], *contextual exploration* goes one step further [43]. Patterns are rules that are fired in two steps: firstly, initial clues (verbs or semantic classes) are searched in sentences and secondly, the contextual clues are searched only in results of the first step. Initial clues characterize the relation type and tend to be stable across domains or textual genre. Contextual clues are expected to be modified to adapt the rule to new corpora and improve its efficiency to identify related terms.

4.3 Pattern Matching

The facilities provided for pattern matching on text determine the quality of the phrases identified as clues of linguistic relations. One of the difficulties when adjusting patterns is to know how they will be matched to text.

Linear Matching*: Each pattern component is searched in sequence and with equal importance.* The majority of pattern-based tools sequentially browse the corpus to find out sentences or phrases that match pattern components. As a consequence, the first components of the pattern are searched more often, even if they do not contribute much to sentence discrimination. So linear matching performs slowly on very large corpora when patterns contain a lot of frequent and non discriminative categories (like prepositions or determiners). But the simplicity of this algorithm makes it quite intuitive for users. It is implemented in Caméléon and Prométhée to browse technical books, as well as in systems where patterns are made of lemma or verbs [39], [12].

Search Restricted to Focused Sentences: One way to reduce the search space when matching patterns to text is to focus on contexts that are more likely to contain linguistic clues of relations. The selection of such contexts relies on other types of knowledge. For instance, RelExt selects sentences where paired terms that often are in collocation cooccur; then it tries to match patterns only with these sentences [26]. In some domains, a semantic or lexical resource can provide pairs of related terms. Patterns are expected to characterize the contexts in which pairs of terms occur in the same sentence. This process is suggested by Hearst to indentify domain specific patterns [25]. New patterns can be learned from these contexts using a supervised learning algorithms [35] or [40] or they can be hand-crafted reading the contexts [19].

Focused Search using Priority Pattern Components: Some search strategies may give a higher priority to pattern components that play a more important role in relation interpretation. When explicitly mentioned, the verb is often considered as one of the strongest contributors to the relation meaning. Matching <Left Verbe-Def Nexus Right> patterns starts by searching the verb Verbe-Dedf; Nexus is used to reject non valid contexts, then Left and Right are tested with regular expressions to look for the related terms. In [9] patterns are matched according to a similar rule: priority is given to the *middle* component, which is expected to convey the relation meaning. In contextual exploration rules [43], the focus is determined by the initial clues. Then contextual clues are searched to identify additional linguistic features that express the terms in relationship [44]. This process accelerates text browsing and identifies complex formulations of relations or variants in these formulations.

4.4 Limitations of pattern-based relation extraction

Many pattern matching implementations do not account for all linguistic phenomena that may arise in language, in particular variation [41]. For instance, the efficiency of patterns is measured with criteria such as precision (ability to

identify valid sentences expressing a semantic relation) and recall (ability to find out all the occurrences of a pattern in the corpus). When trying to improve recall and precision to gain efficiency, patterns become more specific and numerous: each of them accounts for a way of expressing a relation. Such patterns share the same core linguistic clues. The paradox here is that patterns are supposed to be generic and to abstract linguistic phenomena at a higher level. Quantitative evaluations influence what is considered as a good pattern: it is a productive pattern, not one that best accounts for all the linguistic formulations of a relation. For instance, in some systems, learned patterns must have either at least 3 occurrences in the studied corpus or their recall must be higher than a given threshold. Language analysis is secondary while finding pattern utterances is the priority.

4.5 Two alternatives to "hard pattern" that manage flexibility

Learning approaches can improve relation extraction. We report here two alternatives to patterns that reduce some of the above-mentioned limitations.

Soft patterns: The notion of soft pattern contributes to better handle the generality of patterns in real-world corpora like the Web [35]. In GlossExtractor [18], candidates are pruned using more refined stylistic patterns and lexical filters to improve precision while keeping pattern generality. Soft patterns refine this idea thanks to probabilistic lexico-semantic patterns "*that allow a partial matching"* [36]. Instead of a Boolean result, the system calculates a degree of match probability. Soft matching may be carried out in two ways, using either an n-gram language model (Expectation Maximization algorithm), or Profile Hidden Markov Models.

Concept lattices: Learned concept lattices can be an alternative to patterns: links represent hierarchical relations between words, and nodes are clusters of salient words aggregated using synonymy, similarity, or sub-trees of a thesaurus [45]. However, some problems remain, like word selection and aggregation, or word sense disambiguation. The methodology proposed in [45] aligns patterns using of wildcard (*) characters to facilitate sentence clustering. Each cluster of sentences is then generalized to a lattice of word classes. This approach is able to both identify definitions and extract hypernyms. These patterns generalize over lexico-syntactic patterns, and outperform them.

5 Conclusion

Pattern-based relation extraction is one of the most popular ways to identify semantic relations. Although patterns abstract some linguistic regularity, we

identified three types of variation phenomena that influence patterns and their interpretations. We analysed this variation situation to illustrate how it turns relation identification and interpretation into a more complex and subtle task. Confronting the pattern definitions with these variations questions the nature of patterns and even their relevance as a search structure. We raise the question of an alternative to patterns that could better account for variability in semantic relation identification and "understanding".

Pattern-based relation extraction systems have more or less ability to anticipate linguistic and semantic variations. Flexible and adaptable relation definition is desirable to adapt to specific domains. The quality and precision of pattern matching depend on (i) the text pre-processing that is required, (ii) what is searched with the help of a pattern: just a context or a precise triple with a relation label and related terms, or labeled conceptual relations, (iii) the ability to evaluate a confidence degree of each proposed relation.

In short, the most frequent limitations identified for pattern-based approaches are the following: one of the related terms is missing; the pattern is not powerful enough to match with complex variations; there is confusion between arguments in the sentence and concepts in the triple; the estimated POS of some words are wrong; pattern search gives a similar weight to each word in the pattern; patterns are rigid and not adaptable. Relation finding is more efficient and linguistic variation phenomena is better when more complex linguistic variants can be matched to the pattern, and when patterns and target relations can be adapted to the corpus and domain.

From these observations, we stand up for the necessity for linguistics studies in prior to design tools and to integrate human interpretation when representing semantic relations from texts. In terms of research the main questions are: how to design tools that carry out a better linguistic analysis? How to integrate interpretation within the analysis process? How to facilitate the design of ad-hoc patterns? More cross-disciplinary studies (involving corpus linguistic and knowledge engineering) have to be carried out, to identify additional fine-grained linguistic indices for each type of relation. We also expect relation extraction systems to support relation modelling rather than automate it, so that the modelling goal could influence relation interpretation.

References

1. Meyer, I.: Extracting Knowledge–rich Contexts for Terminography: A Conceptual and methodological Framework. In: D. Bourigault, M.C. L'Homme and C. Jacquemin (eds.), Recent Advances in Computational Terminology, pp. 279–302. John Benjamins, Amsterdam (2001)
2. Flowerdew, J.: Definitions in Science Lectures. Applied Linguistics, 13(2), 202–221 (1992)

3. Pearson, J.: The Expression of Definition in Specialized Texts: A Corpus–based Analysis. In: 7th Int. Congress on Lexicography, pp. 817–824 (1996)
4. Marshman, E.: Lexical Knowledge Patterns for the Semi–automatic Extraction of Cause–effect and Association Relations from Medical Texts: A Comparative Analysis of English and French. PhD Dissertation, Département de linguistique et de traduction, Université de Montréal (2007)
5. Condamines A.: Taking genre into account for analyzing conceptual relation patterns, Corpora 8, 115–140 (2008)
6. Garcia, D.: Analyse automatique des textes pour l'organisation causale des actions, Réalisation du système Coatis, PhD. Dissertation in Computer Science, University Paris IV-Sorbonne, (1998).
7. Condamines, A.: Corpus Analysis and Conceptual Relation Patterns. Terminology 8(1), 141–162 (2002)
8. Condamines, A., Rebeyrolle J.: Searching for and Identifying Conceptual Relasionships via a corpus–based approach to a Terminological Knowledge Base (CTKB): method and results. In: D. Bourigault, M.C. L'Homme and C. Jacquemin (eds.), Recent Advances in Computational Terminology, pp. 127–148. Benjamins, Amsterdam/Philadelphia (2001)
9. Agichtein, E., Gravano, L. Snowball: Extracting relations from large plain text collections, In: 5th ACM Conference on Digital Libraries, pp. 85–94. San Antonio, Texas (2000)
10. Barrière C., Aguado A.: TerminoWeb: a software environment for term study in rich contexts, In: Int. Conf. on Terminology, Standardization and Technology Transfer (TSTT 2006), pp. 103–113. Beijing (China) (2006).
11. L'Homme, M.C., Marshman, E.: Terminological Relationships and Corpus–based Methods for Discovering them: An Assessment for Terminographers, In: Bowker, L. (ed.) Lexicography, Terminology, and Translation. Text–based studies in honor of I. Meyer, Univ. of Ottawa Press, pp. 67–80 (2006)
12. Soler V., Alcina A.: Patrones léxicos para la extracción de conceptos vinculados por la relación parte–todo en español, Terminology 14(1), 99–123 (2008)
13. Maedche, A., Staab S.: Discovering conceptual relations from text. In: W. Horn (ed.): ECAI 2000. 14th European Conference on Artificial Intelligence, pp. 321–325, IOS Press, Amsterdam (2000)
14. Aussenac–Gilles, N., Jacques, M.–P.: Designing and Evaluating Patterns for Relation Acquisition from Texts with Caméléon. Terminology 14(1), 145–73 (2008)
15. Buitelaar P., Cimiano P., Magnini B.: Ontology Learning From Text: Methods, Evaluation and Applications, IOS Press, Amsterdam (2005)
16. Malaisé, V., Zweigenbaum, P, Bachimont, B.: Mining defining contexts to help structuring differential ontologies. Terminology, 11(1) 21–53 (2005)

17. Pantel P., Pennachiotti M.: Expresso: leveraging generic patterns for automatically harvesting semantic relations. In: ACL,113–120,Sidney (2006)
18. Velardi, P., Navigli, R., Cuchiarelli, A., Neri, R.: Evaluation of Ontolearn, a methodology for automatic learning of domain ontologies, In: [19] pp. 92–106 (2005)
19. Hearst, M.: Automatic Acquisition of Hyponyms from Large Text Corpora, In: International Conference on Computational Linguistics (COLING–ACL '92), Nantes, pp. 539–545 (1992)
20. Auger, A., Barrière, C.: Pattern based approaches to semantic relation extraction: a state–of–the–art. Terminology 14(1), 1–19 (2008)
21. Halskov, J., Barrière, C.: Web based extraction of semantic relation instances for terminology work. Terminology 14(1), 20–44 (2008)
22. Pantel P., Pennacchiotti, M.: Automatically Harvesting and Ontologizing Semantic Relations. In: P. Buitelaar and P. Cimiano (Eds.) Ontology Learning and Population: Bridging the Gap between Text and Knowledge. pp. 171–198, IOS Press, Amsterdam (2008)
23. Staab, S., Maedche, A.: Ontology Learning for the Semantic Web, IEEE Intelligent Systems, 16(2), 72–79 (2001)
24. Maynard, D., Li, Y., Peter, W.: NLP techniques for Term Extraction and Ontology Population, In: Buitelaar, P., Cimiano P. (Eds.), Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pp. 107–127. IOS Press, Amsterdam (2008)
25. Morin E.: Patrons lexico–syntaxiques pour aider au dépouillement terminologique. Traitement Automatique des Langues, 40–1, 143–166 (1999)
26. Schutz A., Buitelaar P.: RelExt: A tool for relation extraction from text in ontology extension, In: 4th International Semantic Web Conference, Galway, pp. 593–606, Springer Verlag, Berlin, (2005)
27. Barrière, C.: Building a concept hierarchy from corpus analysis, Terminology 10–2, 241–263, (2004)
28. Aussenac-Gilles, N., Despres, S., Szulman, S.: The TERMINAE Method and Platform for Ontology Engineering from texts, In P. Buitelaar, P. Cimiano (Eds.), Bridging the Gap between Text and Knowledge - Selected Contributions to Ontology Learning and Population from Text, pp. 199-223. IOS Press, (2008)
29. Gaizauskas R., Demetriou G., Artymiuk P., Willett P.: Protein structure and information extraction from biological texts : The PASTA system, Bioinformatics 19(1), 135–143 (2003)
30. Mukherjea S., Sahay, S.: Discovering biomedical relations utilizing the world–wide web. In: Pacific Symposium on Bio–Computing. Maui, Hawaii, pp.164–175 (2006)
31. Khelif, K., Dieng–Kuntz, R., Barbry, P.: An ontology–based approach to support text mining and information retrieval in the biological domain, Journal of Universal Computer Science, 13 (12), 1881–1907 (2007)

32. Nédellec C.: Machine Learning for Information Extraction in Genomics – State of the Art and Perspectives, In: Text Mining and its Applications: Results of the NEMIS Launch Conference Series: Studies in Fuzziness and Soft Computing Sirmakessis, pp.99–118. Spiros (2004)
33. Ramakrishnan C., Mendes P., Wang S., Sheth A.: Unsupervised discovery of compound entities for relationship extraction. Lecture Notes in Computer Science, 2008, Vol. 5268/2008, 146–155 (2008)
34. Snow R., Jurafsky D., Ng A.: Learning syntactic patterns for automatic hypernym discovery, NIPS (2005)
35. Navigli, R., Velardi, P., Stefano F.: A Graph–based Algorithm for Inducing Lexical Taxonomies from Scratch. In: 22nd Internat. Joint Conference on Artificial Intelligence (IJCAI 2011), pp. 1872-1877, Barcelona (2011)
36. Girju, R., Badulescu, A., and Moldovan, D.: Automatic Discover of Part-Whole Relations, Computational Linguistics, 32/1, 83-135, (2006)
37. Feliu J.: Relacions conceptuals i terminologia: analisis i proposta de deteccio semiautomatica, Doctoral thesis, Universitat Pompeu Fabra, Barcelona, Espagne (2004)
38. Faure, D., Nedellec, C.: A corpus–based conceptual clustering method for verb frames and ontology, In : LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications, pp. 5–12 (1998)
39. Sierra G., Alarcón R., Aguilar C. , Bach C.: Definitional verbal patterns for semantic relation extraction, In: In: Terminology 14,1, 74–98 (2008)
40. Ruiz–Casado M., Alfonseca E., Castells P.: Automatising the learning of lexical patterns: An application to the enrichment of Wordnet by extracting semantic relationships from Wikipedia, Data and Knowledge Engineering, 61(3), 484–499 (2007)
41. Blohm S., Cimiano P.: Using the Web to Reduce Data Sparseness in Pattern–Based Information Extraction. In: 11th European Conf. on Principles and Practice of Knowledge Discovery in Databases,18–29. Warsaw (2007)
42. Girju, R, Moldovan D.: Text mining for causal relations, FLAIRS 2002, Pensacola Beach (Florida), pp. 360–364 (2002)
43. Desclés J.P., Jouis C., Oh H.G., MaireReppert D.: Exploration contextuelle et sémantique : un système expert qui trouve les valeurs sémantiques des temps de l'indicatif dans un texte, In : Knowledge Modelling and Expertise Transfer (KEMT'91), Amsterdam, pp. 371-400, (1991)
44. Minel J.-L.: Filtrage Sémantique : du résumé automatique à la fouille de textes. Hermès, Paris (2002)
45. Carpineto C. and Romano G.: Using concept lattices for text retrieval and mining. In: Lecture Notes in Computer Science 3626, 161–179. (2005)