# Pattern-based approaches to semantic relation extraction : A state-of-the-art

Auger, Alain; Barrière, Caroline



**Publisher's version / Version de l'éditeur:**
https://doi.org/10.1075/term.14.1.02aug
*Terminology, 14, 1, pp. 1-19, 2008-12-15*

This paper was originally published in Terminology 14:1 (2008), John Benjamins Publishing Co.

# Pattern-based Approaches to Semantic Relation Extraction
## A state-of-the-art

Alain Auger and Caroline Barrière

In recent years, several scientific disciplines such as cognitive science, generative linguistics, artificial intelligence (AI), and computational linguistics have showed growing interest in the many facets of semantic relationships. Some of the representational problems investigated by the AI community in the 1990s (Allen 1995) have found new application grounds with the emerging Semantic Web challenges. Nowadays, several conferences dedicated to specific problems of knowledge acquisition and knowledge representation such as the International Knowledge Capture[1] Conference and the International Semantic Technology Conference,[2] to name a very few, bring together scientists from diverse research communities. For example, in 1997, a workshop entitled *Beyond Word Relations*[3] examined a number of relationship types with significance for information retrieval beyond the conventional topic-matching relationship (Green et al. 2002).

The reader will find good overviews of existing semantic analysis approaches in (Dale et al. 2000)[4] as well as in a two volume study on semantic relations (Bean and Green 2001 and Green et al. 2002). Semantic relations are at the core of any representational system, and are keys to enable next generation of information processing systems with semantic and reasoning capabilities. Acquisition, description, and formalization of semantic relations are fundamental requirements to many natural language processing (NLP) applications.

Semantic networks support the construction and the organization of lexicons, terminologies, taxonomies, and ontologies. Rich sets of semantic relationships have been implemented in well-known projects such as the Unified Medical Language System (UMLS),[5] WordNet,[6] and MultiNet. Multilayered extended semantic networks (abbreviated MultiNet) are both a knowledge representation paradigm and a language for meaning representation of natural language expressions (Helbig 2006). According to

Helbig (2006), MultiNet is one of the most comprehensive and thoroughly described knowledge representation systems. It specifies conceptual structures by means of about 140 predefined relations and functions, which are systematically characterized and underpinned by a formal axiomatic apparatus.[7] As mentioned by Sheth and Lytras "importance of semantics has been recognized in different areas of data and information management, including better access, exchange, interoperability, integration and analysis of data." (Sheth and Lytras 2007: vi)

Automatically extracting semantic relations – the building blocks of ontologies and of any formal knowledge representation system – from textual data is a way of minimizing the labor-intensive phase of manual knowledge engineering and thus overcoming the long-standing knowledge acquisition bottleneck. A comprehensive description of existing ontological engineering methodologies has been presented by Gómez-Pérez et al. (2004). The role of ontologies in natural language processing is discussed in Vossen (2003). The author also presents cognitive, AI, and linguistic traditions to ontological engineering and usage. Specialized ontologies can be seen as the end-product of the terminological tasks of conceptual clarification and knowledge structuring. Several research projects rely on text mining techniques to extract valid semantic relationships from textual datasets in order to generate domain ontologies.[8]

Among different text mining techniques, the pattern-based approaches, pioneered by Hearst (1992), have inspired the work of many and are getting more and more attention in the scientific community. Investigation of automatic ways of finding semantic relations using such approaches is represented by recent work from Cimiano et al. (2005), Pantel and Pennacchiotti (2006), Malaisé et al. (2005), Marshman and L'Homme (2006), Bourigault and Aussenac-Gilles (2003), Auger (1997), to name a few.

A pattern-based approach is a "bottom-up" methodology. It investigates human artifacts such as electronic texts in order to find linguistic means involved in the production and the elicitation of semantic relations. This approach is characterized by two assumptions (a) the target relation is a specific (named) relation and (b) that relation is explicitly expressed in text between words or lexical units.

With respect to (a), it contrasts with approaches which attempt at finding "unnamed" or rather general "similarity" relations between words or terms. Such

approaches (Yu and Agichtein 2003; Dagan et al. 1995; Li and Abe 1998; Lin 1998) are based on clustering methods and follow Harris' distributional hypothesis claiming that words or terms are semantically similar to the extent to which they share similar syntactic contexts. These approaches extend previous work done in automatic thesaurus building (Grefenstette 1994).

With respect to (b), it contrasts with research which attempts to discover the meaning of implicitly expressed relations as found in noun compounds or multi-word expressions (Moldovan et al. 2004; Nastase and Szpakowicz 2003; Rosario and Hearst 2001; Vanderwende 1994). The relation between *laser* and *printer* in *laser printer* is not the same as the relation between *street* and *light* in *street light*. Analysis of syntactic relations as conveyors of semantic relations between lexical units can help structuring a terminology and could certainly be seen in complement to pattern-based expression of relations. Interestingly, a noun-modifier disambiguation task is also presented in a pattern-based study by Turney (2006), with a disambiguation strategy relying on the explicit occurrence in texts of linguistic patterns between the noun and its modifier.

Some approaches aiming at finding both named and explicitly defined semantic relations rely on the resemblance of terms internal structures using morphological analysis (Claveau and L'Homme 2005), and therefore do not assume any external context in which both terms appear. Ibekwe-SanJuan (2006) differentiates "internal evidence" corresponding to morpho-syntactic variations from "contextual evidence" expressed by linguistic patterns in texts. Although the challenges given by the research directions cited above are many and quite interesting, the attention in this special issue of *Terminology* is given to "contextual evidence" of semantic relations.

## Pattern-based Extraction Dimensions

Pattern-based semantic relation extraction frequently involves four main steps: (A) defining the semantic relation of interest, (B) discovering the actual patterns which explicitly express such relation in text as well as the syntactic conditions under which the meaning of the targeted relation is realized, (C) searching for instances of the relation using the patterns, and (D) structuring the new instances as part of a new or existing ontology (or terminological database).

(A) Relations of interest

In information extraction, pattern-based approaches are used to find relations such as *located-in*, *book-authored-by*, *birthdate-of* (Blohm and Cimiano 2007; Ravichandran and Hovy 2002). The work of Alfonseca et al. (2006) explores a multitude of relations using the same general approach, such as *employee-organization*, *painter-painting*, *film-director*, etc. As shown in Malaisé et al. (2005), in terminology, the main relations of interest are those revealing definitional properties of terms. Some relations have been studied much more than others. Among the many studied relations is hypernymy (or *is-a*) (Caraballo 1999; Ravichandran and Hovy 2002), meronymy (or *part-whole*) (Winston et al. 1987; Berland and Charniak 1999; Girju et al. 2003, Pennacchiotti and Pantel 2006), definitional relations (Pasça 2005) and causality (Barrière 2001; Khoo et al. 2002; Girju 2003; Marshman and L'Homme 2006; Pennacchiotti and Pantel 2006). The hypernymy relation has long been at the center of interest since it structures taxonomies and ontologies. Linguistic relations of synonymy and antonymy are also being studied. The distinction between conceptual and linguistic[9] relations is not always taken into account in the literature. They are then grouped under the generic label "semantic relations". Nevertheless, the methods involved in the extraction of conceptual or linguistic relations are generally the same.

An interesting set of relations is tested by Pantel and Pennacchiotti (2006): the traditional *is-a* and *part-whole* relations, as well as *Succession* (e.g. Bush :: Reagan), *Reaction* (magnesium :: oxygen) and *Production* (hydrogen :: metal hydrides). Such a range of relations shows again how pattern-based approaches are both used in factual information extraction and in encyclopedic knowledge extraction.

(B) Patterns

Once relations of interest have been identified, research investigates the linguistic patterns expressing these relations. Research can adopt an onomasiological approach in trying to discover patterns expressing specific relations. Onomasiological methods starts from specific relation such as the Cause-Effect relation and try to identify the linguistic means that can be used to express such a causal relation. Research can also adopt a

semasiological approach where analysis tries to identify which semantic relations can be expressed by specific linguistic markers.

In the context of computational terminology, linguistic markers have been referred to as "knowledge patterns" (KPs) which correspond to the natural language instantiations of semantic relations (Meyer 2001). These KPs help the discovery of useful text utterances, which have been called knowledge-rich contexts (KRCs) (Meyer 2001).

(B1) Discovery

Traditionally, computational lexicography and computational terminology have leveraged on two different types of sources to acquire semantic relations. Existing electronic dictionaries have been used since the 1980's as means to study semantic networks from existing linguistic description of dictionary entries. Véronis and Ide (1991) performed an assessment of semantic information that can be automatically extracted from machine readable dictionaries (MRDs). In fact, a large body of research has been done on the automatic extraction of patterns from MRDs, mostly in the 1980s and the 1990s, before the advent of much available corpus. Typical examples include the work of Richardson et al. (1998) creating MindNet from an encyclopedia and the recent work from Dancette (2007) using encyclopedic articles from the *Analytical Dictionary of Retailing* to extract domain-specific semantic relations. Much of the work done during these years is reviewed by Barrière (2004) and by Sierra et al. (this issue), who refer to work on MRDs as the basis of understanding definitional knowledge.

Nowadays, with the availability of very large textual datasets, corpora are being applied text mining techniques and algorithms to retrieve and describe empirically semantic relations and the contextual lexical units they involve. On of the strategies of pattern-based approaches to relation extraction from textual data consist in compiling lists of reliable patterns that can instantiate specific semantic relation types and use these lists to find new instances in texts to gradually improve the coverage of (existing) ontologies. Such strategies are performed in a cyclic or bootstrapping method. Although Hearst (1992) is cited as an early reference for such technique, more recently Brin (1998) has presented in detail a Dual Iterative Pattern Relation Expansion (DIPRE) approach, demonstrated using the *author-of* relation, but applicable to any relation. Although

usually the seeds of the bootstrapping process consist of a few known pairs of terms instantiating a relation of interest, some other work such as Etzioni et al. (2004) uses a bootstrapping process starting from manually defined trusted patterns. Any bootstrapping approach to semantic relation extraction requires a method to control the expansion phase and avoid drifting. This can be achieved via an automatic assessment of the quality of the new term pairs as well as the quality of the generated patterns. We will discuss this assessment as we further discuss the DIPRE approach in the instance discovery section below.

One important factor in a corpus-based methods is the actual choice of the corpus. As mentioned by Condamines, "the problem of elaborating relational systems from corpora with a linguistic method poses questions about a three-way dependency existing between corpus, relations and patterns." (Condamines 2002: 141) The selection of corpus has a tremendous impact on the results of the knowledge discovery process. For specialized domains, specialized corpora might be used (Morin 1999), and although some approaches have been recently suggested for semi-automatic construction of specialized domain corpora (Barrière and Agbago 2006), such specialized corpora usually remain manually crafted. The problem of data sparseness comes along since specialized corpora are of limited size and the expression of a relation might have a limited number of variations in a specialized dataset. Pattern-based approaches have been criticized in that manner: "The approaches of Hearst and others are characterized by a (relatively) high precision in the sense that the quality of the learned relations is high. However, these approaches suffer from a very low recall which is due to the fact that the patterns are very rare in corpora." (Cimiano et al. 2005: 71)

Exploiting the Internet in order to find patterns has been a recent strategy to cope with the problem of data sparseness in specialized corpora. Nevertheless, with the application of such strategies, recall is boosted and precision decreases. Any hand crafted corpus will tend to be of good quality and will therefore contain limited but reliable knowledge. On the other hand, the Internet contains lots of noisy data. Automatic approaches need to be adapted to the source they work on, and using the Internet forces the focus on increased precision. As reported by Ravichandran and Hovy (2002),

precision varies according to the relation considered. The authors' experiments with specific relations like *birthdates* gave much higher precision results than the *is-a* relation.

Hybrid approaches, such as the ones reported in Blohm and Cimiano (2007) and Pantel and Pennacchiotti (2006), try to balance high reliability of a closed corpus to the high redundancy of the Internet by using different patterns and/or instances to generate filtering strategies which leverage from evidence in both sources. More flavors of these promising hybrid methods are likely to emerge in the near future.

(B2) Pattern Expression

Although linguistic patterns have been called differently by different authors,[10] and the terminology community prefers to refer to them as knowledge patterns, they are frequently referred to as lexico-syntactic patterns. Some research experiments limit the representation of patterns to strings, especially if search on the Internet is involved.[11] Nevertheless, since most of the research has been done so far on closed corpus, patterns are viewed as lexico-syntactic patterns and expressed with a combination of part-of-speech tags and lexical items. For example, a typical hypernymy pattern involving the *is-a* relation would be: NP0 *is-a* NP1 *which* ….

Besides lexical and syntactic characteristics, semantic constraints can also be used to specify patterns. Several approaches involving the use of semantic constraints in patterns or the specification of semantic classes for the terms in relation have been reported in the literature.[12]

(C) Finding instantiations of relations using patterns

In its most basic form, a pattern-based semantic relation would include a term X, a term Y, and a linguistic unit expressing a semantic relation between term X and Y. Finding instances of a semantic relation in texts using linguistic patterns can be implemented in different ways. It can be achieved by building a query where both X and Y are unknown terms linked by a known relation, as for example, *is-a(X,Y)*. Another strategy can be applied to retrieve one unknown parameter and set the second parameter to a known value, for example the pattern *is-a(X,drug)*.

Finding instances is part of a DIPRE bootstrapping process (Brin 1998). During that process, the evaluation of the confidence of patterns and extracted tuples at each iteration is quite important. Only high confidence tuples found in one iteration should be used to find new patterns at the next iteration. In the same way, only high confidence patterns should be used to discover new tuples. This dual constraint leads to methods for measuring pattern confidence and tuple confidence which are interdependent. In the Snowball application (Agichtein and Gravano 2000; Yu and Agichtein 2003), a pattern has higher confidence if it occurs with reliable term pairs, and a term pair is more reliable if it occurs with confident patterns. Such pattern-tuple interdependent reliability estimation is well described in Pantel and Pennacchiotti (2006) "principled reliability measure."

Besides occurring frequency, an important aspect of measuring patterns confidence is their specificity, or their capability at expressing a specific relation and no other relations. This is explored in Alfonseca et al. (2006) who compare their results to a human estimation, and also in Turney (2006) who pushes the notion of specificity further by defining the pertinence of a pattern not with respect to a specific relation but with respect to a specific tuple.

Although much research effort has been invested by different authors on pattern and tuple evaluation, much research remains to be performed in this area as it is a crucial part of the success of the bootstrapping methods to semantic relation extraction.

(D) Knowledge Structuring

The structuring of the knowledge using instances extracted from text is another important task in knowledge formalization. One can use standards such as RDF or OWL[13] to properly formalize and structure conceptual classes, instances and their relationships. Implementation will face typical problems of efficiency, scalability, and reusability.

Existing ontological resources such as DOLCE,[14] SUMO,[15] OpenCyc[16] and the Basic Formal Ontology[17] (BFO) can be used either in supervised approaches to find instances of semantic relations or they can be used as a target reference model to structure and formalize new instances of semantic relations. These ontological resources can also be used to infer new knowledge from facts contained in texts.

Evaluation

We have already mentioned evaluation as being an essential and integral part of the extraction process, especially for guiding the expansion phase as in the DIPRE process (Brin 1998). In terminology work, patterns are usually manually defined and their intrinsic evaluation is not performed. They are evaluated indirectly by the quality of the instances they can retrieve.[18]

Automatic evaluation of the performance of an application at retrieving instances of semantic relation requires the development of gold standards. Defining gold standards requires human judges to manually evaluate and annotate datasets containing instances of semantic relationships. Such gold standards allow the comparison of different systems using typical measures of precision and recall. Table 1 below shows, for a few applications, the actual task to be performed and the gold standard used or developed by the authors for evaluating the task.

Table 1 – Different tasks and evaluation methods (ordered by year of publication)

| **Reference** | **Corpus (used for discovery)** | **Use of external sources** | **Task** | **Gold standard / Evaluation** | **Measure** |
|---|---|---|---|---|---|
| Blohm and Cimiano 2007 | Wikipedia | Internet for pattern generation | Find tuples for set relations | List of tuples semi-automatically built via Wikipedia Categories (albumBy, bornInYear, currencyOf, headquarteredIn, locatedIn, productOf, teamOf) | Precision Recall |
| Pantel and Pennacchiotti 2006 | TREC-9 (5M words newswire) / CHEM (300K words college chemistry textbook) | Internet for pattern search and instance validation | Find tuples for set relations | Five relations: two general (is-a, part-of), one in TREC-9 (succession), two in CHEM (reaction, production) – Random sample of instances evaluated by 2 human judges. | Precision Relative recall (defined by authors) |
| Alfonseca et al. 2006 | Internet | Named Entity Recognition (NER) module | Find a better precision estimator for patterns | Two human annotators<br>19 relations (death year, soccer team/city, director/film, etc.) | Precision |
| Turney 2006 (exp. 1) | N/A (not a discovery task) | Waterloo Multitext application (find patterns) | Answer the analogy questions | 374 college-level multiple-choice word analogies (SAT tests) | Score on test |
| Turney 2006 (exp. 2) | N/A (not a discovery task) | Waterloo Multitext application (find patterns) | Noun-Modifier classification | 600 manually labelled noun-modifier pairs from (Nastase and Szpakowicz 2003) | Precision Recall |

| | | | | | |
|---|---|---|---|---|---|
| Greenwood and Stevenson 2006 // Stevenson and Greenwood 2005 | MUC-6 | No | Find documents and sentences | MUC-6 Corpus of annotated documents and sentences (within the documents) for their pertinence about different movements of executives in companies (appointed-by, promoted-by, works-for, resigns-from) | Precision Recall |
| Etzioni et al. 2004 | Internet | No | NER | 5 classes: City, USState, Country (found in the Tipster Gazetteer) Actor, Film (found in the Internet Movie Database) | Precision Recall |
| Cimiano et al. 2005 | Collection of texts from two tourist-related sites | (a) Internet for pattern generation and instance validation (b) Wordnet for instance validation | Test a set of discovered *isa(X,Y)* | Ontology about tourism hand-made by an ontology engineer with 289 concepts (only is-a links) | Precision Recall |
| Yu and Agichtein 2003 | 52000 scientific journal articles | Gene taggers | Find gene synonyms | Gene synonyms extracted from SWISSPROT and judged by six biology experts (for recall) Sampling of 200 synonymy pairs evaluated by 2 biology experts (for precision). | Precision Recall |
| Agichtein and Gravano 2000 | 300000 newspaper documents | No | Finding at least one occurrence of a relation | 13000 Organizations found on Hoover's Online website Relation of Organization-Location | Precision Recall |
| Moldovan et al. 2000 | Internet | No | Wordnet expansion | User validation of new concepts for seeds in the financial domain | Yes/No (equivalent to Precision) |
| Brin 1998 | 147 GB (24M web pages) Stanford WebBase | No | Finding instances of the relation | Author-Title / Manual comparison of 20 generated books picked randomly to Amazon directory | Yes/No (equivalent to Precision) |

The work of Marshman and L'Homme (2006) and Barrière (2001) discuss pattern evaluation issues in a terminological context.

Knowledge discovery techniques applied to ontological engineering can use existing ontologies as gold standards to train and test new knowledge discovery algorithms and to try to automatically derive the same ontology from domain-specific texts.[19] Table 1 shows an example with the work of Cimiano et al. (2005). This task is facing the challenge of measuring and comparing the quality of empirical textual data against subjectively built ontologies representing subject matter experts' views and interpretations of their knowledge domain. Even if subjectivity was not a concern, the automatic comparison of extracted knowledge to already existing knowledge in the

ontology often requires sophisticated natural language processing tools to take into consideration different types of variations (lemmatization, terminological variants, etc.).

## Terminological issues

Although much work discussed so far is not applied to terminology, the pattern-based semantic relation extraction approaches involved are basically the same as the ones used in computational terminology. As mentioned earlier, terminology work (Grabar and Hamon 2004) has focused more on relations of hypernymy, hyponymy, synonymy, meronymy, holonymy, function, and causality which are important in defining terms and their relationships. Computational terminology is interested in the semantic relation patterns themselves, in understanding, describing, and formalizing their linguistic properties, and in analyzing them beyond their discovery capability.

Many relations such as hypernymy and meronymy are not domain-specific. Other relations can be observed mainly within a specific domain. For example, the biomedical domain uses specific types of causal relations (Marshman and L'Homme 2006; Rosario and Hearst 2004).

In terminology, there is a practical aspect in using patterns, that is to help the terminologist, or knowledge worker, finding definitional information in text. For example, different applications such as OntoLearn (Navigli et al. 2003), CAMÉLÉON (Séguéla and Aussenac-Gilles 1999), TerminoWeb (Barrière and Agbago 2006) or Corpógrafo (Sarmento et al. 2006) all include an important user interaction aspect. The latter three mentioned also integrate manually defined linguistic patterns respectively for French, English and 5 different languages (mainly Portuguese, but also English, Spanish, Italian and French).

## Contributions to this special issue

This special issue contains five different contributions, exploring a large spectrum of questions related to pattern-based approaches to semantic relations extraction.

The first contribution, by Halskov and Barrière, takes on the challenge of pattern discovery and instance discovery in a biomedical domain. Their research addresses the difficult problem of evaluation and ranking of discovered knowledge. Although, a large

terminology database already exists for biomedicine, that of UMLS, their work clearly shows how it can be extended with success using the Internet to search for new instances via automatically discovered linguistic patterns. They evaluate their approach against a human established gold standard for four relations, that of synonymy, hypernymy, *may-prevent* and *induces*. Their approach is close in spirit with lots of DIPRE style work found in the information extraction community. Their interest in terminology, as opposed to general language, makes them refine their instance filtering with specialized language heuristics, such as measuring the "termhood" of their newly extracted instances using a comparative corpus approach.

The next three contributions introduce knowledge patterns and evaluate their performance at discovering of new instances. This is typical to terminology work, where manual exploration is very valuable to provide a deeper understanding of the factors impacting the discovery, coverage, and productivity aspects of linguistic patterns.

As part of the understanding of the human effort involved in the development of a specialized-domain ontology, Aussenac-Gilles and Jacques explore the problem of domain-dependency and ask the pertinent question of the level of human effort needed to reuse linguistic patterns from another domain. They revisit the notion of "generic" versus "specific" patterns as frequently mentioned in the literature. They suggest "reusable" as a more appropriate way of describing patterns which transpose well from one domain to another. Their work is on the French language, which, although certainly not studied as much as English, has been studied by a few other researchers (Séguéla and Aussenac-Gilles 1999; Marshman et al. 2002; Malaisé et al. 2004; Claveau and L'Homme, 2004). Their system, CAMÉLÉON, is quite representative of development efforts in computational terminology which focuses on providing terminologists and knowledge workers with good interactive support for tasks such as taxonomy and ontology building.

Besides English and French, in depth investigations in other languages is not frequent in the literature. Although Corpógrafo (Sarmento et al. 2006) does include patterns for five languages, including Spanish, not much theoretical or extensive pattern development work has been presented in the literature for Spanish. It is therefore very interesting to present two studies on this language in this special issue.

Sierra and co-authors introduce and describe a small set of manually designed knowledge patterns for Spanish. They propose a syntactic approach both to represent patterns and to design instance filtering heuristics. Their focus is on verbal patterns and the particularities of the different types of definitional contexts they introduce: analytical, extensional, functional or synonymic contexts.

Soler and Alcina, also working on Spanish, introduce and evaluate knowledge patterns for a set of relations and explore more specifically the *part-whole* relation.

Adopting a different angle, this issue's last contribution by Marshman identifies and investigates one of the most challenging issue in the extraction of valid and fully qualified semantic relations. This issue is the one of the level of certainty of relations expressed in texts. The author demonstrates how languages such as French and English use several linguistic means to express the level of certainty of a given relation. Adverbs such as *likely*, *probably*, *possibly* are few of the several means that can be used to determine certainty of the existence of a given relation. The description and formalization of certainty / uncertainty is crucial in ontology building and knowledge representation. Without means to cope with uncertainty, ontologies and taxonomies will assume that facts extracted from texts all meet the same truth condition. It will therefore not be possible to run inference engines and to properly exploit the knowledge contained in ontologies.

## Concluding Remarks

The field of pattern-based approaches to semantic relations extraction is currently very active. The new semantic relation classification task at SemEval 2007 (Girju et al. 2007) is another sign of renewed interest in this area, interest which had started in the late 1980s with much work on machine readable dictionaries (MRDs). With electronic texts now largely available, most pattern-based work has moved from MRDs to corpus, and even recently to the Internet and to very large datasets such as the Terabyte Corpus at TREC. Huge heterogeneous datasets certainly bring their own idiosyncrasies, questions and problems.

This paper was originally published in Terminology 14:1 (2008), John Benjamins Publishing Co.

The main challenges pertaining to semantic relation extraction have been summarized by Pantel and Pennacchiotti (2006).

> The following desiderata outline the properties of an ideal relation harvesting algorithm:
> - Performance: it must generate both high precision and high recall relation instances;
> - Minimal supervision: it must require little or no human intervention;
> - Breadth: it must be applicable to varying corpus sizes and domains; and
> - Generality: it must be applicable to a wide variety of relations (i.e. not just is-a or part-of).

This introduction does not give complete answers to these challenges but rather presents the general problem areas of pattern-based semantic relation extraction and more specifically emphasize the challenging task of discovering these linguistic patterns in text.

This special issue takes a terminology stance and provides a state of the art account of the type of terminological work currently dedicated to the study and use of pattern-based approaches to support semantic relation extraction tasks.

Aknowledgments

The authors would like to thank Marie-Claude L'Homme and Annaïch Le Serrec for reviewing and commenting a previous version of this introduction.

The authors would like to thank the members of the Program Committee for this Special Issue: Sophia Ananiadou (University of Manchester), Nathalie Aussenac-Gilles (IRIT, CNRS), Lynne Bowker (Université d'Ottawa), Vincent Claveau (IRISA, Rennes), Anne Condamines (Université Toulouse-le-Mirail), Béatrice Daille (LINA, Université de Nantes), Patrick Drouin (OLST, Université de Montréal), Fidelia Ibekwe-SanJuan (Université de Lyon 3), Kyo Kageura (University of Tokyo), Elizabeth Marshman (OLST, Université de Montréal), Roberto Navigli (University of Rome "La Sapienza"), Maria Teresa Pazienza (AI Research Group, University of Rome "Tor Vergata"), Pascale Sébillot (IRIRA, Rennes), Pierre Zweigenbaum (LIMSI, Paris)

This paper was originally published in Terminology 14:1 (2008), John Benjamins Publishing Co.

## Notes

[1] http://www.k-cap.org
[2] http://www.semantic-conference.com/
[3] Hetzler, B. (1997). Beyond word relations. SIGIR Forum, 31/2, 28-33.
[4] In particular, chap. 5 on symbolic approaches to semantic analysis and chap. 24 on empirical approaches to lexical knowledge acquisition
[5] http://www.nlm.nih.gov/research/umls/
[6] http://wordnet.princeton.edu/
[7] http://en.wikipedia.org/wiki/MultiNet; http://pi7.fernuni-hagen.de/forschung/multinet/multinet_en.html
[8] For instance, the SACOT project in Canada is dealing with the application of NLP techniques in support to the semi-automatic construction of ontologies from texts. Similar projects and frameworks are currently being implemented. See http://www.cs.utexas.edu/users/mfkb/related.html for a list of worldwide projects related to ontologies.
[9] A relation occurring between terms rather then between concepts.
[10] See Halskov and Barrière, this issue
[11] Search engines do not allow for part-of-speech to be used. They do not have proximity operators allowing for strings to be found close to each other.
[12] See for instance: Agichtein and Gravano (2000), Marshman and L'Homme (2006), Stevenson and Greenwood (2005), Yu and Agichtein (2003), Greenwood and Stevenson (2006)
[13] Semantic Web standards available at http://www.w3c.org
[14] Descriptive Ontology for Linguistic and Cognitive Engineering. (http://www.loa-cnr.it/DOLCE.html)
[15] The Suggested Upper Merged Ontology (http://www.ontologyportal.org/)
[16] http://www.opencyc.org
[17] Basic Formal Ontology project (http://www.ifomis.uni-saarland.de/bfo/)
[18] Such approach is applied in three contributions to this special issue, that of Aussenac-Gilles & Jacques, Sierra et al., and Soler & Alcina.
[19] This is one of the many challenges being addressed by the SACOT research project at Defence R&D Canada. For more information, contact Alain.Auger@drdc-rddc.gc.ca

## References

Agichtein, E. and L. Gravano. 2000. "Snowball: Extracting relations from large plaintext collections." In *Proceedings of the 5th ACM International Conference on Digital Libraries*. 85-94. San Antonio, Texas.

Alfonseca, E., N. Ruiz-Casado, M. Okumura and P. Castells. 2006. "Towards large-scale non-taxonomic relation extraction: Estimating the precision of rote extractors." In *Proceedings of the 2nd Workshop on Ontology Learning and Population*. 49-56. Sydney, Australia.

Allen, J. 1995. *Natural Language Understanding*. Redwood City, CA: Benjamin/Cummins Publishing Company.

Auger, A. 1997. *Repérage des énoncés d'intérêt définitoire dans les bases de données textuelles*. Thèse de doctorat. Université de Neuchâtel. (http://doc.rero.ch/search.py?recid=473&ln=fr). Accessed November 5, 2007 .

Barrière, C. 2001. "Investigating the causal relation." *Terminology* 7(2): 135-154.

Barrière, C. 2004. "Knowledge-rich contexts discovery." In *Seventeenth Canadian Conference on Artificial Intelligence (AI'2004)*. 187-201. London, Canada.

Barrière, C. and A. Agbago. 2006. "TerminoWeb: A Software Environment for Term Study in Rich Contexts." In *International Conference on Terminology, Standardisation and Technology Transfer (TSTT'2006)*. 103-113. Beijing, China.

Bean, C. A. and R. Green. 2001. *Relationships in the Organization of Knowledge*. Dordrecht; Boston: Kluwer Academic Press.

Berland, M. and E. Charniak. 1999. "Finding parts in very large corpora." In *Proceedings of ACL-1999*. 57-64. College Park, Maryland.

Blohm, S. and P. Cimiano. 2007. "Using the Web to reduce data sparseness in pattern-based information extraction." In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases*. 18-29. Warsaw, Poland.

Bourigault, D. and N. Aussenac-Gilles. 2003. "Construction d'ontologies à partir de textes." In *Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues* (TALN 2003). 27-50. Batz-sur-Mer, France.

Brin, S. 1998. "Extracting patterns and relations from the World Wide Web." In *The World Wide Web and Databases, International Workshop WebDB'98*. 172-183. Valencia, Spain.

Caraballo, S. 1999. "Automatic acquisition of a hypernym-labeled noun hierarchy from text." In *Proceedings of ACL-99*. 120-126. College Park, Maryland.

Cimiano, P., A. Pivk, L. Schmidt-Thieme and S. Staab. 2005. "Learning taxonomic relations from heterogeneous sources of evidence." In Paul Buitelaar, Philipp Cimiano and Bernardo Magnini, *Ontology Learning from Text: Methods, evaluation and applications*. 55-73. Amsterdam: IOS Press.

Claveau, V. and M.-C. L'Homme. 2004. "Discovering specific semantic relationships between nouns and verbs in a specialized French corpus." In the *3rd Edition of CompuTerm Workshop (CompuTerm'2004) at Coling'2004*. 39-46. Geneva, Switzerland.

Claveau, V. and M.-C. L'Homme. 2005. "Structuring terminology using analogy-based machine learning." In *Proceedings of TKE'2005* (Terminology and Knowledge Engineering). PAGES. Copenhagen, Denmark.

Condamines, A. 2002. "Corpus analysis and conceptual relation patterns." *Terminology* 8(1): 141-162.

Dagan, I., S. Marcus and S. Markovitch. 1993. "Contextual word similarity and estimation from sparse data." *Proceedings of ACL-1993*. 164-171. Columbus, Ohio.

Dale, R., H. Moisl and H. Somers. 2000. *Handbook of Natural Language Processing*. New York. Marcel Dekker Inc.

Dancette, J. 2007. "Semantic Relations in the Field of Retailing." *Terminology* 13(2): 201-233.

Etzioni, O., M. Cafarella, D. Downey, S. Kok, A.-M. Popescu, T. Shaked, S. Soderland, D.S. Weld and A. Yates. 2004. "Web-scale information extraction in KnowItAll." *Proceedings of the 13th International World Wide Web Conference*. 100-110. New York, New York.

Girju, R. 2003. "Automatic detection of causal relations for question answering", in *Proceedings of ACL Workshop on Multilingual Summarization and Question Answering*, 76-83. Sapporo, Japan.

Girju, R., A. Badulescu and D. Moldovan. 2003. "Learning semantic constraints for the automatic discovery of part-whole relations." in *Proceedings of HLT/NAACL-03*. 80-87, Edmonton, Canada.

Girju, R., P. Nakov, V. Nastase, S. Szpakowicz, P. Turney and D. Yuret. 2007. "SemEval-2007 Task 04: Classification of Semantic Relations between Nominals." In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, 13-18, Prague, Czech Republic.

Gómez-Pérez, A., M. Fernández-López and O. Corcho. 2004. *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. London: Springer

Grabar, N. and T. Hamon. 2004. "Les relations dans les terminologies structures : de la théorie à la pratique." *Revue d'Intelligence Artificielle (RIA)* 18(1): 57-85.

Green, R., C. Bean and S.H. Myaeng. 2002. *The Semantics of Relationships. An interdisciplinary perspective*. Dordrecht, The Netherlands: Kluwer Academic Press.

Greenwood, M.A. and M. Stevenson. 2006. "Improving semi-supervised acquisition of relation extraction patterns." In *Proceedings of the Workshop on Information Extraction Beyond The Document*. 29-35. Sydney, Australia.

Grefenstette, G. 1994. *Explorations in Automatic Thesaurus Discovery*. Boston, MA: Kluwer Academic Publisher.

Hearst, M.A. 1992. "Automatic acquisition of hyponyms from large text corpora." In *Proceedings of COLING-92*. 539-545. Nantes, France.

Helbig, H. 2006. *Knowledge Representation and the Semantics of Natural Language*. Berlin/Heidelberg/New York: Springer.

Hetzler, B. 1997. "Beyond word relations". *SIGIR Forum*, 31(2): 28-33.

Ibekwe-SanJuan, F. 2006. "Clustering semantic relations for constructing and maintaining knowledge organization tools." *Journal of Documentation* 62(2): 229-250.

Khoo, C., S. Chan and Y. Niu. 2002. "The many facets of the cause-effect relation." In Green, R., C. Bean and S.H. Myaeng (eds.). *The Semantics of Relationships. An interdisciplinary perspective*. 51-70. Dordrecht, The Netherlands. Kluwer Academic Press.

Li, H. and N. Abe. 1998. "Word clustering and disambiguation based on co-occurrence data." In *Proceedings of Coling-ACL 1998*. 749-755. Montreal, Canada.

Lin, D. 1998. "Automatic retrieval and clustering of similar words." In *Proceedings of Coling-ACL 1998*. . 768-774. Montreal, Canada.

Maedche, A. and S. Staab. 2000. "Mining non-taxonomic conceptual relations from text." In *Knowledge Engineering and Knowledge Management. Methods, models, and tools: 12th International Conference, EKAW 2000. Proceedings*. 189-202. Berlin: Springer.

Malaisé, V., P. Zweigenbaum and B. Bachimont. 2005. "Mining defining contexts to help structuring differential ontologies." *Terminology* 11(1): 21-53.

Malaisé, V., P. Zweigenbaum and B. Bachimont. 2004. "Detecting semantic relations between terms in definitions." In the *3rd Edition of CompuTerm Workshop (CompuTerm'2004) at Coling'2004*. 55-62. Geneva, Switzerland.

Marshman, E. and M.-C. L'Homme. 2006. "Disambiguation of lexical markers of cause and effect." In Picht, H. (ed.). *Modern Approaches to Terminological Theories and Applications*. *Proceedings of the 15th European Symposium on Language for Special Purposes, LSP 2005*. 261-285. Bern: Peter Lang.

Marshman, E., T. Morgan and I. Meyer. 2002. "French patterns for expressing concept relations." *Terminology* 8(1): 1-29.

Meyer, I. 2001. "Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework." In Bourigault, D., C. Jacquemin and M.-C. L'Homme (eds.). *Recent Advances in Computational Terminology*. 279–302. Amsterdam/Philadelphia: John Benjamins.

Moldovan, D., R. Girju and R. Rus. 2000. "Domain-specific knowledge acquisition from text." In *Proceedings of the 6th Conference on Applied Natural Language Processing*. 268-275. Seattle, WA.

Moldovan, D., A. Badulescu, M. Tatu, D. Antohe and R. Girju. 2004. "Models for the semantic classification of noun phrases." In *Proceedings of the Computational Lexical Semantics Workshop at HLT-NAACL* 2004. 60-67. Boston, MA.

Morin, E. 1999. "Automatic acquisition of semantic relations between terms from technical corpora." *Proceesings of the 5th Internationall Congress Terminology and Knowledge Extraction* (TKE-99). 268-278. Vienna: TermNet.

Nastase, V. and S. Szpakowicz. 2003. "Exploring noun-modifier semantic relations." In *Fifth International Workshop on Computational Semantics (IWCS-5)*. 285-301. Tilburg, Netherlands.

Navigli, R., P. Velardi and A. Gangemi. 2003. "Ontology learning and its application to automated terminology translation." *IEEE Intelligent Systems*: 18(1). 22-31. New York.

Pantel, P. and M. Pennacchiotti. 2006. "Espresso: Leveraging generic patterns for automatically harvesting semantic relations." In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*. 113-120. Sydney, Australia.

Pasça, M. 2005. ".Finding instance names and alternative glosses on the Web: WordNet reloaded." In *CICLing 2005*, LNCS 3406. 280-292. Berlin/Heidelberg: Springer-Verlag

Pennacchiotti, M. and P. Pantel. 2006. "Ontologizing semantic relations." In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 793-800. Sydney, Australia.

Ravichandran, D. and E.H. Hovy. 2002. "Learning surface text patterns for a question answering system." In *Proceedings of ACL-2002*. 41-47. Philadelphia, Pensylvania.

Richardson S.D., Dolan W.B. and L. Vanderwende. 1998. "MindNet: Acquiring and structuring semantic information from text." In *Proceedings of ACL-Coling* 1998. 1098-1102. Montreal, Canada.

Rosario, B. and M. Hearst. 2001 "Classifying the semantic relations in noun-compounds via a domain specific hierarchy." In *2001 Conference on Empirical Methods in Natural Language Processing*. 82-90. Ithaca, New York.

Rosario, B. and M. Hearst. 2004. "Classifying semantic relations in bioscience text." In *ACL'04*. 430-437. Barcelona.

Sarmento, L., B. Maia, D. Santos, A. Pinto and L. Cabral. 2006. "Corpógrafo V3 – From terminological aid to semi-automatic knowledge engineering." In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*. 1502-1505. Genoa, Italy.

Séguéla, P. and N. Aussenac-Gilles. 1999. "Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine." In *Actes de la conférence Ingénierie des Connaissances* (IC'99). 79-88. Palaiseau, France.

Sheth, A. and M. Lytras. 2007. *Semantic Web-based information systems. State-of-the art applications*. London, UK: Cybertech Publishing.

Stevenson, M. and M.A. Greenwood. 2005. "A semantic approach to IE pattern induction." In *Proceedings of the 43rd Annual Meeting of the ACL*. 379-386. Ann Arbor, MI.

Turney, P. 2006. "Expressing implicit semantic relations without supervision." In *Proceedings of the 21rst International Committee on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics* (COLING/ACL 2006).313-320. Sydney, Australia.

Vanderwende, L. 1994. "Algorithm for automatic interpretation of noun sequences." In *Proceedings of 15th ACL*. 782-288. Las Cruces, NM.

Véronis, J. and N. Ide. 1991. "An assessment of semantic information automatically extracted from machine readable dictionaries." In *Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*. 227-232. Berlin, Germany.

Vossen, P. 2003 "Ontologies", In Ruslan Mitkov (ed.). *The Oxford Handbook of Computational Linguistics*. 464-482. OUP, Oxford.

Winston, M., R. Chaffin and D. Hermann. 1987. "A taxonomy of part-whole relations." *Cognitive Science* 11: 417-444.

Yu, H. and E. Agichtein. 2003. "Extracting synonymous gene and protein terms from biological literature." *Bioinformatics* 19(1), 340-349.