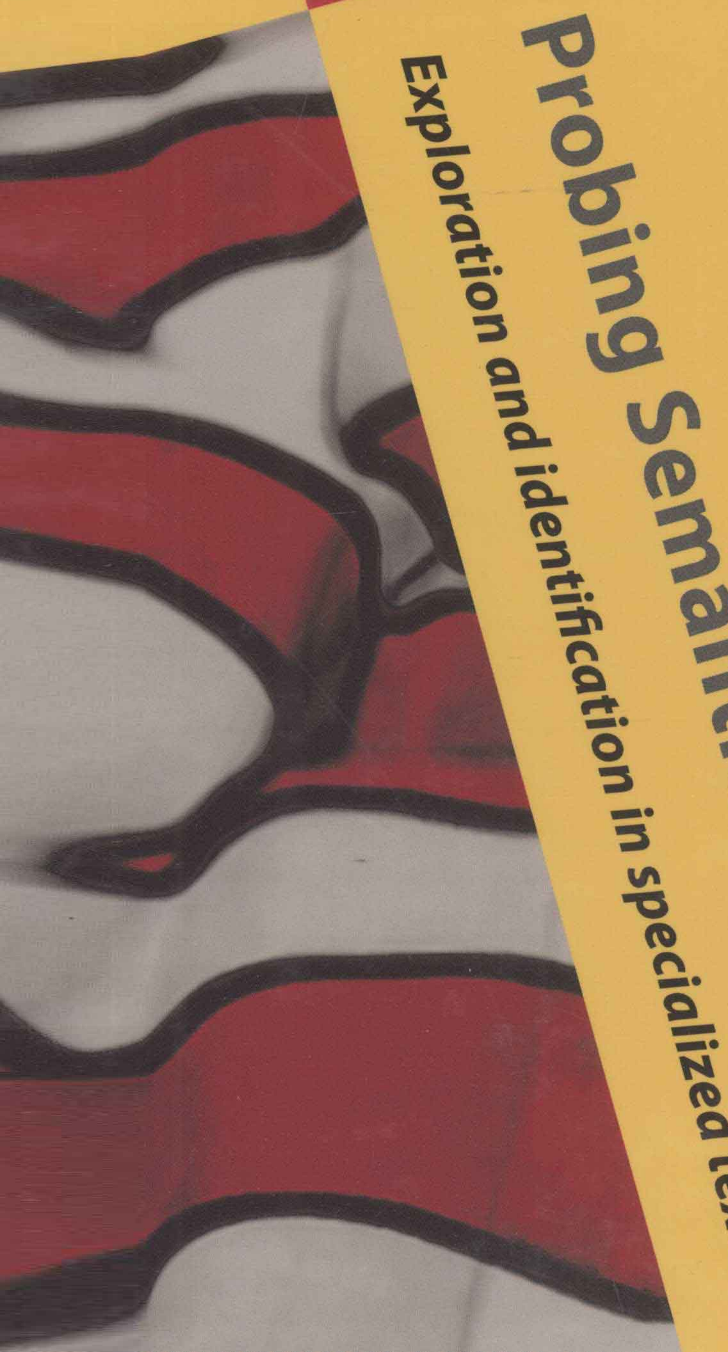
Alain Auger and
Caroline Barrière (eds.)

Probing Semantic Relations

Exploration and identification in specialized texts



BENJAMINS CURRENT TOPICS

23

Probing Semantic Relations

Exploration and identification in specialized texts

Edited by

Alain Auger
Defence Research and Development Canada (Valcartier)

Caroline Barrière
National Research Council of Canada

John Benjamins Publishing Company
Amsterdam / Philadelphia

™ The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI z39.48-1984.

Library of Congress Cataloging-in-Publication Data

Probing semantic relations : exploration and identification in specialized texts / edited by Alain Auger, Caroline Barrière.

p. cm. (Benjamins Current Topics, ISSN 1874-0081 ; v. 23)

Includes bibliographical references and index.

1. Semantics. 2. Relational grammar. I. Auger, Alain. II. Barrière, Caroline.

P325.P7223 2010

401'.43--dc22 2010022885

ISBN 978 90 272 2253 4 (Hb ; alk. paper)

ISBN 978 90 272 8792 2 (Eb)

John Benjamins Publishing Co. · P.O. Box 36224 · 1020 ME Amsterdam · The Netherlands
John Benjamins North America · P.O. Box 27519 · Philadelphia PA 19118-0519 · USA

About the Authors

Alain Auger received a Ph.D. degree in Linguistics from Université de Neuchâtel in Switzerland in 1997. He has developed new methods to capture semantic information in electronic texts. His main research interests include knowledge capture and representation, terminology extraction, text and audio mining, ontology engineering, machine translation, automated document summarization and automated document classification. He joined Defence Research and Development Canada (DRDC) as a Defence Scientist in 2003. He is leading a group of scientists dedicated to Research and Development in the field of Intelligence Production and Exploitation. Most of his work is currently being applied to the development of new capabilities for the Intelligence community. He is the project manager and lead scientist on several research projects at DRDC.

Caroline Barrière has a doctorate in Computational Linguistics from Simon Fraser University, as well as a master's degree in Electrical Engineering and a bachelor's degree in Computer Engineering from École Polytechnique de Montréal. She worked as Assistant Professor at University of Ottawa's School of Information and Technology Engineering (SITE) from 1997 to 2003, and then became a research officer at the Interactive Language Technology Group of the National Research Council of Canada. Her work focuses on computational terminology and lexicography. She is particularly interested in the automatic extraction of knowledge from dictionaries and corpora, as well as the conceptual representation of this knowledge. She further aims at applying her research in the development of tools for terminologists, translators and language teachers.

Jakob Halskov has a doctorate in Computational Linguistics from Copenhagen Business School and a master's degree in English from Copenhagen University: He has worked as research assistant at the Danish Language Council since 2007. His main research interests are corpus linguistics and computational terminology. He is particularly interested in the task of automatic knowledge extraction from natural language text, but also in the semantic fuzziness which arises when non-experts make use of terminology from specialized fields in their everyday communication.

Nathalie Aussenac-Gilles graduated in 1986 from an engineering school in computer science and obtained her Ph.D. in 1989. She became a CNRS (French National Research Agency) researcher in 1991. Since then, she is a member of the IRIT laboratory in Computer Science at University "Paul Sabatier" of Toulouse. Her research interests include knowledge engineering, natural language processing and terminology based approaches for ontology engineering from text. Her work is influenced by long term cross-disciplinary collaborations with linguists and researchers in human factors.

Marie-Paule Jacques obtained her Ph.D. in Linguistics in 2003. After temporary positions as assistant professor at Toulouse le Mirail University and post-doc researcher at IRIT (Paul Sabatier University, Toulouse) and at LIPN (Laboratoire d'Informatique de Paris-Nord), Paris 13

University, she is lecturer at Marc Bloch University, Strasbourg, since 2007. Her research interests focus on corpus linguistics, Natural Language Processing (NLP) and linguistic studies that aim at improving NLP (such as discourse organisation, anaphora processing, tagging and parsing), terminology and retrieval of conceptual relations from texts. She has been involved in projects that implied collaborations with other linguists, with researchers in informatics and with psychologists.

Gerardo Sierra is a full-time researcher of the Engineering Institute of National Autonomous University of Mexico (UNAM). He is the Leader of the Language Engineering Group, founded in 2000. He has received his Ph.D. in Computational Linguistics at UMIST, Manchester, in 1999. He has taught courses of Introduction to Language Engineering, Corpus Linguistics and Text Mining at the UNAM and other universities. His research interests are in the area of Language Engineering, specifically Computational Lexicography, Terminotics, Corpus Linguistics, Semantic Relations and Text Mining. He has published some important papers oriented to Language Engineering in the *International Journal of Lexicography*, *Terminology* and *Lecture Notes in Computer Science.*

Rodrigo Alarcón is a Ph.D. student in Language Sciences at the Institut Universitari de Lingüística Aplicada of UPF, Barcelona, since 2003. He has received a Diploma in Advanced Studies in Applied Linguistics in 2006. His Ph.D. Research Dissertation focuses on developing a definitional contexts extraction system from specialised corpora in Spanish. His research interests are Corpus Linguistics, Information Extraction, Text Mining, and Language Engineering.

César Aguilar is a Ph.D. student in Linguistics student in the Department of Linguistics of UNAM, Mexico City, since 2003. He has received a Masters Degree on Hispanic Linguistics from the UNAM in 2003. His Ph.D. Research Dissertation focuses on the description and analysis of definitional verbal patterns associated with definitions in definitional contexts in Spanish. His research interests are Computational Lexicology, Formal Grammars for Automatic Parsing and Language Engineering.

Carme Bach is a professor of Catalan Linguistics at University Pompeu Fabra and a researcher of the consolidated research group IULATERM at Pompeu Fabra University (Barcelona, Spain). Her research focuses on General and Specialised Discourse Analysis, Lexicography and Corpus Linguistics. Her publications deal with connectives, reformulation and its importance in the process of specialised discourse construction and in extraction of semantic information of specialised discourse. Her Ph.D. thesis, which is about reformulation markers, includes a lexicographical implementation prototype for these units.

Victoria Soler lectures Terminology and Translation Technologies in the Translation and Interpreting degree and in the master's degree on Translation Technologies and Localization at the Universitat Jaume I of Castellón, Spain. She is currently completing her Ph.D. thesis on the automatic extraction of conceptual relations in a specialized corpus. She is also an active member of the research group TecnoLeTTra.

Amparo Alcina is a Senior Lecturer at the Universitat Jaume I of Castellón (Spain) where she teaches Translation Technologies and Terminology to translators. She is also the director of the Translation and Interpreting degree, and the master's degree on Translation Technologies and Localization. Her research interests are in translation and language technologies, and terminology. She coordinates the research group TecnoLeTTra (http://tecnolettra.uji.es), and leads the research project ONTODIC that aims to create onomasiological dictionaries based on ontologies.

Elizabeth Marshman is an Assistant Professor at the University of Ottawa School of Translation and Interpretation and a member of the Observatoire de linguistique Sens-Texte (OLST). She received her doctorate in Translation (Terminology option) from the Département de linguistique et de traduction of the Université de Montréal in 2007, and also worked as a research assistant at the OLST. Her thesis research focused on the interlinguistic comparison of knowledge patterns in English and French. Elizabeth obtained her M.A. and B.A. in translation from the University of Ottawa (1997, 2002). Her research interests include computer-assisted tools for terminology, terminography and translation.

Table of contents

Probing semantic relations

Exploration and identification in specialized texts

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 the 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 multiword 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 volume 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 start 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 volume), 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. One 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 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." (2002: 141) The selection of a 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 corpora, 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 of 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 automatic extraction process, especially for guiding the expansion phase as in the DIPRE process (Brin 1998). In terminology work, linguistic patterns are often 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.

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.

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
Turney 2008	N/A (not a discovery task	Waterloo Multitext application (find patterns)	Questions about synonymy, antonymy, and associations	320 TOEFL synonym questions, 136 ESL synonym/antonym questions, 144 word pairs classified similar/associated/both	Accuracy (score on tests)
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 Recal
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 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

Reference	Corpus (used for discovery)	Use of external sources	Task	Gold standard / Evaluation	Measure
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 *is-a(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 SWISS-PROT 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)