

L-ISA: Learning Domain Specific Isa-Relations from the Web

Alessandra Potrich, Emanuele Pianta

Fondazione Bruno Kessler
38050 Povo (Trento), Italy
potrich@fbk.eu, pianta@fbk.eu

Abstract

Automated extraction of ontological knowledge from text corpora is a relevant task in Natural Language Processing. In this paper, we focus on the problem of finding hypernyms for relevant concepts in a specific domain (e.g. Optical Recording) in the context of a concrete and challenging application scenario (patent processing). To this end information available on the Web is exploited. The extraction method includes four main steps. Firstly, the Google search engine is exploited to retrieve possible instances of isa-patterns reported in the literature. Then, the returned snippets are filtered on the basis of lexico-syntactic criteria (e.g. the candidate hypernym must be expressed as a noun phrase without complex modifiers). In a further filtering step, only candidate hypernyms compatible with the target domain are kept. Finally a candidate ranking mechanism is applied to select one hypernym as output of the algorithm. The extraction method was evaluated on 100 concepts of the Optical Recording domain. Moreover, the reliability of isa-patterns reported in the literature as predictors of isa-relations was assessed by manually evaluating the template instances remaining after lexico-syntactic filtering, for 3 concepts of the same domain. While more extensive testing is needed the method appears promising especially for its portability across different domains.

1. Introduction

Domain ontologies are widely considered as crucial ingredients for many applications, and especially so for the Semantic Web. However, building large coverage domain ontologies is difficult and expensive, since it usually demands for the intervention of experts in the specific domain. On the other hand, the request for domain ontologies is growing and so is the need for developing new methodologies for the automatic acquisition of ontological knowledge from texts.

Studies on automatic acquisition of ontological knowledge can be classified in two main groups: those related to the acquisition of new concepts and relations between them (e.g., a device is an artifact) broadly classified as *ontology learning*, and those related to the acquisition of factual knowledge about specific instances (e.g., Einstein is an instance of a scientist; Einstein was born in 1879), known as *ontology population*. This paper will focus on the ontology learning task.

In the last years, a number of Natural Language Processing techniques have been developed which try to automatically extract concepts and their relations from corpora. Among relations the most frequently targeted is the isa-relation. A widely used technique for the acquisition of this relation, requires the identification of linguistic patterns (Hearst, 1992; Hearst, 1998) expressing the relation in texts. A number of such patterns has been proposed in the literature (Hearst, 1992; Hearst, 1998; Mititelu, 2006); yet, with the exclusion of (Snow et al., 2005), little is reported about systematic assessments of pattern reliability as predictors of the relation. Even less is known about the applicability of such techniques to domain specific ontologies in concrete real world contexts.

In this paper we report on an ontology learning technique based on the Web and how it can be exploited to derive ontological knowledge for a specific domain, in a concrete, challenging application scenario (that is patent semantic

analysis). We will also provide an evaluation of the contribution of the most commonly used patterns to 1) the acquisition of isa-relation, 2) the solution of our specific problem.

The present work is developed as part of the PatExpert European project, whose goals demand the representation of the relevant content of patents in a knowledge base. In order to achieve this aim, relation extraction techniques are employed which require the definition in the ontology of the most important concepts occurring in the patents. Even for a limited domain as, for example, that of Optical Recording, this implies the definition of thousands of concepts in the knowledge base.

2. Building the Optical Recording Domain Ontology

The PatExpert Optical Recording Domain Ontology (ORDO) has been built in three stages. Firstly, 200 concepts were manually defined in the ontology and possibly linked to more general concepts in SUMO and to equivalent synsets in WordNet. The selection of this first group of concepts was based on the list of the most frequent terms automatically extracted from the corpus.

Then, a first ontology learning approach (Pro-ISA) has been used which exploits the knowledge encoded in WordNet and WORDNET DOMAINS¹. Following this approach, fragments of the WordNet hierarchy are projected on the ORDO ontology, thereby defining a novel part of the ontology (AUTO-ORDO).

Consider, for example, the term “cross-talk”; this lexical unit belongs to one Wordnet synset (i.e. the term is non ambiguous), but the synset is not linked to any concept in the ORDO ontology. Thus no relation involving instances of “cross-talk” can be added to the Knowledge Base, unless a new domain concept is added to ORDO in correspondence

¹the labeling of WordNet synsets with domain information.

with the WordNet synset. For this purpose, Pro-ISA looks for the hypernyms of the {cross-talk} synset in search of a synset having the following two properties: 1) its domain is compatible with the Optical Recording domain, based on WORDNET DOMAINS labeling²; and 2) it is linked to a concept of the ontology. Let's suppose this is the {event} synset. Then, Pro-ISA can induce that the chain of isa-relations going from {cross-talk} to {event} corresponds to a chain of isa-relations in the ontology, and that such chain can be projected on the AUTO-ORDO ontology.

Unfortunately, this first learning strategy works only when a term expressing a relevant concept in the target domain is defined as a concept in WordNet. This is not the case in around 15% of cases. To solve the remaining cases we devised a second learning technique (L-ISA) which creates a new concept for the term not present in WordNet as hyponym of some concept defined either in ORDO or in WordNet.

To extract this new isa-relation we initially considered the idea of exploiting the patents' corpus, but we discarded it later, for two main reasons: 1) in limited corpora this relation is not frequent (Caraballo, 1999); and 2) in the patents' domain, it is not uncommon that new definitions for terms are given with a local scope. As an alternative, we decided to extract isa-relations using lexico-syntactic patterns from the richest resource of specialized information available: the Web (Sombatsrisomboon R. and M., 2003).

3. Extracting isa-relations from the Web

Given a term which is relevant for the target domain (e.g. because of its frequency in a reference corpus) and which cannot be automatically linked to ORDO using the first learning strategy, L-ISA looks in the Web for textual patterns representing the isa-relation and containing the relevant term.

The patterns reported in the literature as expressing the isa-relation have usually the structure "NP1 isa-phrase NP2", where NP1 and NP2 are noun phrase constituents expressing a hyponym and its hypernym, while *isa-phrase* is a sequence of tokens. In fact, in our scenario the hyponym term is known in advance, and the patterns become therefore more specific: "TERM-NP isa-phrase HYPER-NP" or "HYPER-NP isa-phrase TERM-NP".

These are partly syntactic patterns that cannot be used directly for searching the Web, because search engines as Google cannot directly recognize syntactic structures. Thus, given a lexico-syntactic pattern to be searched we proceed in two steps: first we derive a more general token-based pattern and submit it to Google retaining the first 100 snippets (at most). Then, we analyze the snippets with NLP tools and derive more specific lexico-syntactic patterns from them.

3.1. Snippet Acquisition

Let's consider the following example. Given: 1) the term "photodetector", which we cannot assign to ORDO, and 2)

the pattern "TERM-NP is an HYPER-NP", we build the string query "photodetector is an" and submit it to Google. Here follows a sample snippet returned by Google.

"... upper frequencies, the PIN waveguide **photodetector** is an attractive device, since it is possible to reduce transit time without .."

Returned snippets are HTML-parsed and transformed in pure text.

3.2. Lexico-syntactic filtering

Snippets are then annotated with TextPro – a suite of NLP tools developed at FBK-Irst. Each token is annotated with PoS, lemma and chunk information. This allows for identifying the two NP chunks containing the target term and its candidate hypernym. More precisely, we look for the pattern (B-NP? I-NP*) where B-NP marks the beginning of a noun phrase and I-NP marks the rest of the NP (see Tab. 1).

	token	PoS	lemma	chunk
TERM-NP	the	AT0	the	B-NP
	PIN	NN1	pin	I-NP
	waveguide	NN1	waveguide	I-NP
	photodetector	NN1	photodetector	I-NP
isa-phrase	is	VBZ	be	B-VP
	an	AT0	an	B-NP
HYPER-NP	attractive	AJ0	attractive	I-NP
	device	NN1	device	I-NP

Table 1: Instance of isa-pattern with TextPro annotation.

The extracted NP chunks need further filtering:

TERM-NP: we filter out NPs which include noun modifiers of the target term (see for instance "PIN waveguide photodetector" above) or which look like proper names (e.g., uppercase letter in the middle of a sentence).

HYPER-NP: only a restricted number of PoS-patterns are kept:

(N | AN | NN | NNN | ANN | XNN | R Vpastpart | AXN)

where, according to WordNet, A stands for adjective, N for noun, R for adverb, Vpastpart for verb in past participle form, and X means unknown to WordNet.

Table 2 reports some TERM-NP/HYPER-NP couples after lexico-syntactic filtering.

TERM-NP	HYPER-NP
the photodetector	analog signal
A photodetector	apparatus
photodetector	effective monitor
The photodetector	electric device
a photodetector	electronic device
photodetector	object

Table 2: Isa instances after lexico-syntactic filtering.

²Each WordNet Domain label was manually given a value expressing the level of compatibility with the Optical Recording domain – namely no, low, midlow, mid, high.

3.3. Semantic filtering and best candidate selection

The lexico-syntactic filtering is followed by a semantic filtering in order to keep only those HYPER-NPs compatible with the Optical Recording domain. This is done by checking whether the HYPER-NP is already a label in one of the known ontologies (SUMO, ORDO, AUTO-ORDO) or, in case it is not, whether it is present in a WordNet synset with a WORDNET DOMAIN label compatible with the Optical Recording domain. Table 3 reports the results of semantic filtering for the *photodetector* example.

HYPER-NP	inKB	inWN	Dom.Comp.
analog signal			
<i>apparatus</i>	yes	yes	midlow
effective monitor			
electric device			
<i>electronic device</i>	yes		
<i>object</i>	yes	yes	midlow

Table 3: The columns “InKB”, “inWN” and “Dom.Comp” specify respectively whether the candidate HYPER-NP of *photodetector* is present in ORDO, in WordNet and the compatibility of the WORDNET DOMAIN labels with the domain of Optical Recording.

The HYPER-NPs that pass this step are all candidates to be used as hypernyms of the target term and hyponyms of some already defined concepts. However, we are interested in selecting only one of these concepts, which is done applying a suitable heuristic. More precisely, each candidate hypernym is given a weight taking into account the frequency and the reliability of the patterns in which it appears (see Tabs. 5, 6), whether it appears in different patterns, and its belonging to specific ontologies (ORDO, AUTO-ORDO or SUMO, in decreasing preference order). Finally, the algorithm selects as hypernym the candidate with the higher weight. A threshold is set to avoid the selection of candidate hypernyms supported by insufficient evidence. In case no candidate is found inside ORDO, AUTOORDO or SUMO we consider the candidates present in WordNet. This happens, for example, for the terms *defocus*, *erasing* and *warpage* (see Tabs. 7, 8).

4. Evaluation

In order to evaluate the accuracy of the the L-ISA algorithm, two kinds of evaluations have been carried out. The first evaluation aims at assessing the reliability of the patterns reported in the literature as predictors of the isa-relation. The second aims at measuring the accuracy of the algorithm in finding the hypernym of a given domain concept.

For the first evaluation we gathered a list of patterns reported in the literature as expressing the isa-relation (Hearst, 1992; Mititelu, 2006). The complete list of significant patterns³ is reported in the first column of Tabs. 5, 6. Then, we singled out 3 representative items from the list of terms that we could not link to the ORDO ontology:

³For the definition of significant pattern see the caption of Tab. 5.

“groove”, “photodetector” and “magnetic head”. These are all domain specific terms.

The isa extraction technique described in section 3 was used to derive about 9,000 snippets for the three target terms. Out of these, about 1,450 instances of isa-relations passed the lexico-syntactic filtering (see Tab. 4). and form the Gold Standard for the pattern reliability evaluation.

	HTML snippets	snippets without patents	lexico-syntactic filtering
groove	5441	5029	764
magnetic head	1209	799	230
photodetector	2353	1934	454

Table 4: For each term, the number of snippets returned by Google (first column), the number of snippets after removing the patents’ snippets (second column), and the number of patterns instances after the lexico-syntactic filtering (third column) are reported. Figures in third column represent the number of evaluation items in the Gold Standard.

Each pattern instance has been assigned to one of three classes, depending on whether it contains: 1) an isa-relation; 2) a relation which is not an isa-relation; 3) an uncertain relation.

The classification of each pattern instance was done trying to understand the intentions of the author. The annotator would assign an instance to class 1 only if he/she thinks that the author intended to say that the class of a certain object (referred to by the target term) is a subclass of some other class (referred to by the hypernym). The evaluation process brought to our attention a non negligible number of cases where linguistic constructions which are usually associated to the isa-relation have in fact a different semantics, expressing the use, or function or subjective perspective on some object class, more than its inclusion in some more general ontological class. For instance if someone writes “to look at details such as landscape, which has a large influence on wind speeds”, he/she does not implies that landscapes belong to the ontological class of details. The implication here is more likely that the landscape is one of the possible objects on which a human agent can focus his/her attention.

The results of this first evaluation are summarized in Tabs. 5, 6, where for each pattern and term, the number of snippets returned by Google, the number of annotated pattern instances and the number of patterns actually expressing an isa-relation are reported. In the last column the average precision is shown. Note that these data are exploited in the final step of the L-ISA extraction algorithm to weight the contribution of the various template instances to the confidence score associated to each candidate hypernym.

The second evaluation was carried out considering the most frequent 100 terms that we were not able to link to the ORDO ontology using the Pro-ISA learning strategy. Note that some of the terms should not be considered at all as relevant terms (see for instance *comprised*). These terms are included in the list because of errors in the linguistic

pattern	groove			magnetic head			photodetector			mean
	HTML	tot	prec	HTML	tot	prec	HTML	tot	prec	
HYPO is a kind of HYPE	8	1	100	0	0	-	4	0	-	100
HYPO is a type of HYPE	4	0	-	1	0	-	0	0	-	-
HYPO is a HYPE	100	38	2.6	23	3	66.6	100	30	30	33
HYPO is an HYPE	100	30	0	14	2	50	66	17	47	32.3
HYPOs are HYPE	100	3	0	100	5	40	100	7	57.1	32.3
HYPE such as HYPO	100	1	0	28	8	62.5	36	3	0	20.8
HYPE such as HYPOs	100	43	23.2	67	38	65.7	100	43	74.4	54.4
HYPE such as the HYPO	100	9	11.1	18	5	40	21	2	50	33.7
HYPE such as a HYPO	100	35	5.7	31	5	80	48	10	100	61.9
such HYPE as HYPO	100	1	0	28	8	50	36	3	0	16.6
such HYPE as HYPOs	100	43	23.2	67	38	65.7	100	43	74.4	54.4
HYPE including HYPO	100	3	33.3	26	4	0	81	5	0	11.1
HYPE including HYPOs	100	36	16.6	35	16	81.2	100	29	41.3	46.3
HYPE including the HYPO	100	5	0	17	2	0	27	3	33.3	11.1
HYPE including the HYPOs	66	22	13.6	8	0	-	14	1	0	6.8
HYPE especially HYPO	100	0	-	3	1	0	0	0	-	0
HYPE especially HYPOs	48	3	0	2	0	-	26	1	0	0
HYPE especially the HYPO	57	8	37.5	2	0	-	3	0	-	37.5
HYPE especially the HYPOs	25	11	18.1	2	1	100	2	0	-	59
HYPE e.g. HYPO	100	1	0	10	0	-	31	4	75	37.5
HYPE e.g. HYPOs	100	23	26	13	1	100	43	8	50	58.6
HYPE in particular HYPO	36	0	-	2	0	-	6	0	-	-
HYPE in particular HYPOs	35	5	20	3	0	-	9	0	-	20
HYPE particularly the HYPO	31	7	14.2	6	0	-	5	0	-	14.2
HYPE particularly the HYPOs	17	2	100	1	0	-	2	1	100	100
HYPE particularly HYPO	79	0	-	5	1	0	3	0	-	0
HYPE particularly HYPOs	27	0	-	7	0	-	9	1	0	0
HYPE except HYPO	100	4	25	0	0	-	2	0	-	25
HYPE except HYPOs	19	0	-	3	0	-	3	1	0	0
HYPE as HYPO	100	0	-	55	1	0	100	5	0	0
HYPE as HYPOs	100	8	12.5	84	7	14.2	100	14	14.2	13.6
HYPE notably HYPO	18	0	-	0	0	-	0	0	-	-
HYPE notably HYPOs	2	0	-	1	0	-	0	0	-	-
HYPE usually HYPO	100	0	-	0	0	-	0	0	-	-
HYPE usually HYPOs	65	1	0	1	0	-	5	0	-	0
HYPE mostly HYPO	100	3	0	0	0	-	2	1	0	0
HYPE mostly HYPOs	33	2	0	0	0	-	1	0	-	0
HYPE mainly HYPO	63	0	-	0	0	-	0	0	-	-
HYPE mainly HYPOs	41	0	-	0	0	-	2	0	-	-

Table 5: Estimates of patterns' reliability as predictors of the isa-relation (part I). The following patterns return 0 snippets for all 3 terms and are not included in the table: "HYPO in common with other HYPE", "HYPOs in common with other HYPE", "HYPO and sometimes other HYPE", "HYPOs and sometimes other HYPE", "HYPO or other kind of HYPE", "HYPOs or other kind of HYPE", "HYPO is a special type of HYPE".

analysis phase. We included these terms in the evaluation to check if, given a "wrong" target term, the algorithm does not find any plausible hypernym (desirable behaviour).

For each term, L-ISA extracted the best candidate hypernym using the method explained above (see Sec. 3.3.). In Tabs. 7, 8 the best candidate for each term is shown. If the first candidate is wrong (column third) the second or third candidate (if correct) is reported in column fourth. Out of 100 terms, 44 gave no result because the information extracted was insufficient to make a decision. In the remaining 56 cases, 12 errors occurred; for 4 of them the second candidate was a correct answer. This preliminary

evaluation suggests that the algorithm can reach an accuracy of 78.6%. Note also that for at least some of the cases in which the algorithm did not provide any answer (*comprised, incremented, limitative, therebetween* ...) this was indeed the expected and desirable behaviour.

5. Related Work

Many works have considered the problem of automatically building or extending an ontology starting from Hearst (Hearst, 1992; Hearst, 1998) who first proposed to gather from texts the syntactic patterns specific to a given relation. A similar approach is employed in (Girju et al., 2006) for

pattern	groove			magnetic head			photodetector			mean
	HTML	tot	prec	HTML	tot	prec	HTML	tot	prec	
HYPE like HYPO	100	7	0	13	1	0	26	0	-	0
HYPE like HYPOs	100	7	0	15	8	50	41	17	82.3	44.1
HYPO like other HYPE	17	3	0	0	0	-	1	0	-	0
HYPOs like other HYPE	7	1	0	0	0	-	2	0	-	0
HYPO as a HYPE	100	7	0	18	2	0	100	14	7.1	2.3
HYPO as an HYPE	100	15	6.6	9	0	-	41	7	0	3.3
HYPE even HYPOs	100	0	-	1	0	-	8	0	-	-
HYPO as well as the HYPE	100	13	0	8	0	-	24	4	0	0
HYPOs as well as the HYPE	100	17	0	6	1	0	22	5	0	0
HYPE other than the HYPO	35	5	0	7	1	0	9	0	-	0
HYPE other than the HYPOs	23	3	0	2	0	-	3	1	0	0
HYPE not least HYPO	6	0	-	0	0	-	0	0	-	-
HYPE but not HYPO	19	0	-	0	0	-	0	0	-	-
HYPE but not HYPOs	10	3	0	0	0	-	1	1	0	0
HYPE for example HYPO	90	1	0	7	0	-	12	0	-	0
HYPE for example the HYPO	63	1	0	13	0	-	19	0	-	0
HYPE i.e. HYPO	100	5	20	10	0	-	22	3	33.3	26.6
HYPO another HYPE	100	6	0	7	0	-	26	1	100	50
HYPO an HYPE	100	4	0	23	2	0	100	8	0	0
HYPO a HYPE	100	2	0	60	9	0	100	10	10	3.3
HYPO a kind of HYPE	14	2	0	0	0	-	0	0	-	0
HYPOs a kind of HYPE	100	7	0	0	0	-	0	0	-	0
HYPE called HYPO	100	0	-	8	0	-	6	0	-	-
HYPE called HYPOs	100	18	55.5	5	0	-	13	3	66.6	61
HYPO and other HYPE	100	9	0	16	6	66.6	31	7	57.1	41.2
HYPOs and other HYPE	100	30	56.6	48	21	61.9	100	22	90.9	69.8
HYPO and many other HYPE	24	4	0	0	0	-	1	0	-	0
HYPOs and many other HYPE	9	0	-	0	0	-	4	2	50	50
HYPO and in other HYPE	6	1	0	4	0	-	0	0	-	0
HYPOs and in other HYPE	7	2	0	1	0	-	0	0	-	0
HYPE or HYPO	100	33	9	79	14	14.2	100	34	5.8	9.6
HYPE or HYPOs	100	68	8.8	34	9	0	100	37	8.1	5.6
HYPO or other HYPE	100	26	19.2	10	0	-	18	3	100	59.6
HYPOs or other HYPE	100	47	4.2	13	4	100	13	3	66.6	56.9
HYPO or any other HYPE	40	6	16.6	3	0	-	9	1	0	8.3
HYPOs or any other HYPE	40	10	10	3	0	-	6	1	100	55
HYPO is HYPE	100	5	0	100	3	66.6	100	5	0	22.2
HYPO is another HYPE	100	16	0	1	0	-	5	2	0	0
HYPO is the HYPE	100	22	9	20	3	33.3	100	30	6.6	16.3
HYPO is the only HYPE	46	10	0	1	0	-	1	1	0	0
HYPO is any HYPE	11	0	-	1	0	-	2	0	-	-

Table 6: Estimates of patterns' reliability as predictors of the isa-relation (part II).

the part-of relation. Other authors propose different approaches. For example, in (K. Shinzato, 2005) the HTML Tags of itemization are employed.

(Snow et al., 2005) use Minipar to save and generalize the contexts (dependency-paths) where an isa-relation occurs. With this method the authors can compare their results with those obtained using a subset of the patterns proposed by Hearst.

Finally, the authors in (Sombatsrisomboon R. and M., 2003) propose to extract the hypernym/hyponym of a concept from the snippets returned by Google, where the query is "X is a" (to find the hypernym of X) and the query "is a Y" (to find the hyponym of Y).

6. Conclusions and future work

The method proposed here for the automatic acquisition of isa-relation appears promising, although a test on a wider set of terms is needed for a more secure assessment of the method's qualities. A point of strength for this method is its portability across different domains. Indeed, a change of domain requires only to redefine the compatibility of WORDNET DOMAINS labels with the new domain. Less straightforward is the portability to other languages because tools and resources for the target language (e.g. PoS tagger, chunker, WordNet, isa-patterns) would be needed. Finally, machine learning methods could be applied to the optimiza-

tion of the weights associated to the patterns.

7. Acknowledgments

This work was carried out with the financial support of the PatExpert European Project.

8. References

- Sharon A. Caraballo. 1999. Automatic construction of a hypernym-labeled noun hierarchy from text. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 120–126, Morristown, NJ, USA. Association for Computational Linguistics.
- Roxana Girju, Adriana Badulescu, and Dan I. Moldovan. 2006. Automatic discovery of part-whole relations. *Computational Linguistics*, 32(1):83–135.
- Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING*, pages 539–545.
- Marti A. Hearst, 1998. *WordNet: An Electronic Lexical Database*, chapter Automated discovery of wordnet relations. MIT Press, Cambridge, MA.
- K. Torisawa K. Shinzato. 2005. Automatic acquisition of hyponymy relations from html documents. *Journal of Natural Language Processing*, 12(1):125–150.
- Verginica Barbu Mititelu. 2006. Automatic extraction of patterns displaying hyponym-hypernym co-occurrence from corpora. In *First Central European Student Conference in Linguistics*.
- Rion Snow, Daniel Jurafsky, and Andrew Y. Ng. 2005. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems 17*, pages 1297–1304. MIT Press, Cambridge, MA.
- Matsuo Y. Sombatsrisomboon R. and Ishizuka M. 2003. Acquisition of hypernyms and hyponyms from the www. In *Proceedings of the 2d International Workshop on Active Mining*.

term	L-ISA result	eval	alternative result
aberration compensator			
amorphization			
axial runout			
barcode	ORDO:information	+	
baseplate	ORDO:component	+	
BER	ORDO:measure	+	
binarized			
cd rom	ORDO:material	-	ORDO:medium
CD-RW	ORDO:cd	+	
centreline	ORDO:line	+	
circumferentially			
coercivity	ORDO:property	+	
colorant	ORDO:material	+	
comparator	ORDO:circuit	+	
compensator	SUMO:device	+	
comprised			
cross-erasing			
curing	SUMO:process	+	
cyclability			
decentering	SUMO:process	+	
defocus	WNhigh:aberration	+	
delamination	AUTOORDO:type	-	
demodulated			
depth dp			
depth gd			
DSP	ORDO:system	+	
DSV			
DVD-RW	ORDO:disc	+	
DVD-Video	ORDO:disc	+	
encoder	SUMO:device	+	
energization			
erasability			
erasing	WNhigh:operation	+	
forming	SUMO:process	+	
HDD	ORDO:component	+	
identifier	ORDO:data	+	
imaged			
incremented			
irradiating			
lambda max			
lambda rms			
length nt			
limitative			
line a-a			
line b-b			
loci	ORDO:region	+	
LPF	ORDO:filter	+	
LPP	ORDO:system	-	
mechanical isolator			
method of cryptocommunication			

Table 7: Post-hoc evaluation of the L-ISA algorithm (part I). Column 2 reports the best L-ISA candidate which is manually evaluated in column 3. If the first result is wrong (“-”), we report in column 4 an alternative right result if ranked by L-ISA as second or third.

term	L-ISA result	eval	alternative result
micrograph	ORDO:data	-	AUTOORDO:photograph
microstructure	ORDO:property	+	
mmol			
MPU	SUMO:device	+	
MTF	ORDO:measure	+	
nsec			
objective lens ol			
oligomer	ORDO:component	-	AUTOORDO:compound
overview	ORDO:information	+	
phase comparator			
photocoupler			
photoresist	ORDO:material	+	
photosensor	SUMO:device	+	
plaintext	ORDO:information	+	
PMA	SUMO:agent	-	
polarizer	ORDO:filter	+	
polycarbonate	ORDO:material	+	
polymethylmethacrylate	ORDO:material	+	
postambles			
power pr			
preamplifier	ORDO:component	+	
preformat			
pregroove			
PSN	AUTOORDO:network	-	
pulse section op			
radii	ORDO:property	+	
readout	AUTOORDO:feature	-	AUTOORDO:instrument
recordable			
recrystallization	SUMO:process	+	
rms	ORDO:measure		
satisfying tmax			
SIL	ORDO:measure	-	
SNR	ORDO:measure	+	
spacer	SUMO:device	+	
spectra	ORDO:data	+	
subcode	ORDO:data	+	
substituent			
substituents			
subtractor			
superposed			
superposition	ORDO:property	-	
surface roughness ra			
TDMA	ORDO:system	-	
therebetween			
tilting	AUTOORDO:action	+	
TOC	ORDO:location	+	
transmissive			
transmissivity	ORDO:measure	+	
VCM	AUTOORDO:product	-	
warpage	WNhigh:distortion	+	

Table 8: *Post-hoc evaluation of the L-ISA algorithm (part II).*