Building a specialized ontology: Why go on the web?
Halskov, J.; Barrière, Caroline

Proceedings of the 8th International Conference on Terminologyand Knowledge Engineering (TKE 2008): "Managing Ontologies and Lexcial Resources", August 2008, Copenhagen, Denmark, 2008

National Research Council Canada
Institute for Information Technology

Conseil national de recherches Canada
Institut de technologie de l'information

*Building a Specialized Ontology: Why Go on the Web? **

Halskov, J., and Barrière, C.
August 2008

* Proceedings of the 8th International Conference on Terminology and Knowledge Engineering (TKE 2008): "Managing Ontologies and Lexcial Resources". August 2008. Copenhagen, Denmark. pp. 175-190.

Building a specialized ontology: Why go on the web?

Jakob Halskov and Caroline Barrière

This research proposes a comparison of two sources of information for building a specialized ontology: the WWW, a large repository of uncategorized texts, and BioMed, a small specialized corpus in the medical domain. The methodology explored is the use of knowledge patterns. These are explicit markers in text leading to semantic or conceptual relations. Although the method developed has interest for discovering new information in order to enrich the UMLS (a biomedical metathesaurus), we measure its success by an attempt to "rediscover" information already present in the UMLS Metathesaurus. Measures of precision and recall are used in several experiments of instance retrieval for four semantic relations important in the UMLS Methathesaurus, two of a general nature (is-a, synonymy) and two domain specific ones (preventing, inducing). Results show that although the WWW is a noisy repository, its exploration has potential and does allow the discovery of valuable specialized knowledge.

1. Introduction

(Hearst, 1992) originally presented a way of identifying semantic relation instances in text by employing what has later been called "knowledge patterns" (Meyer 2001: p290), ”knowledge probes” (Ahmad 1992) or ”explicit relation markers” (Bowden 1996). In this article we will adopt Meyer's terminology and use the acronym KP for such patterns. As Table 1.1 illustrates, useful KPs may occur in either the left, middle or right context of the term pair.

Table 1.1: Knowledge patterns in context

Left	**Term1**	Middle	**Term2**	Right
<causes of	diarrhea	**include>**	parasites	, some cancers , and other…
	diarrhea	**<induced by>**	bacteria	is a typical ...
to minimize the	stomach irritation		aspirin	**<can cause>**
a **<side effect of**	nicotinic acid	**is>**	flushing	

Of course, KPs can also be discontinuous and span more than one of these contexts. We focus our exploration on the middle context following results of (Agichtein, 2000) which show that it is the most informative for English.

Multiple pattern-based relation extraction systems have been developed since (Hearst, 1992), but many systems are custom-tailored to handle specific relation types and use, for example, domain-specific filtering techniques (e.g. Yu, 2002), domain-specific lexicons (e.g. Gaizauskas, 2003) or Named Entity Recognizers (e.g. Mukherjea, 2006). Also, most systems operate on domain-specific corpora like MEDLINE abstracts. Contrarily, a system called WWW2REL, developed by (Halskov and Barrière, 2008) uses the WWW both for discovering KPs and relation instances and can be applied to any relation type.

In the present work, we explore the relation of Synonymy, Is-a, May-Prevent and Induces which are of importance in the UMLS ontology, and for which the WWW2REL system has been used to discover, from the WWW, four sets of KPs (see Halskov and Barrière, 2008). We use these sets of discovered KPs to find instances of the relations both on the WWW and in the BioMed corpus. We then compare the instances retrieved to the ones found in the UMLS ontology which serves as our gold standard for our experiments. We wish to explore to what extent using the web as a specialized corpus affects both recall and precision.

Section 2 briefly presents the KP discovery and highlights possible noise problems when using KPs to discover relation instances. Section 3 describes our experiments and the data. Section 4 presents some comparative experiments on BioMed and WWW. Section 5 presents further experiments for the WWW, and section 6 summarizes our conclusions.

2. Using KPs for instance retrieval

Table 2.1 shows the sets of KPs for the four relations explored. These KPs were discovered automatically in the WWW2REL system using term pairs from the UMLS Metathesaurus and finding recurrent middle contexts.

All details of the methodology for automatically discovering KPs are described in (Halskov and Barrière, 2008). Noise reduction strategies were implemented to obtain a list of "good" KPs, that is KPs which tend to occur with multiple term pairs (general within a relation) and which tend to only occur with term pairs instantiating the particular relation they express (specific to one relation). Even with such strategies Table 2.1 reveals how some KPs are more reliable than others. As no human intervention was allowed during the KP discovery process, this noise remains.

Table 2.1 – Discovered KPs for 4 relations

Relation type	KPs
INDUCES	(1) cause, may cause, to cause, does not cause, which causes, will cause, causing, can also cause, without causing, that can cause, causes, can cause (2) induce, induces, to induce, which induces, or induce, will induce, can induce, for inducing, may induce, induced (3) produce, produces, to produce, does not produce, produced (4) can lead to, leading to, may lead to, leads to (5) overdose include, poisoning include, poisoning (6) promotes, can result in
MAY_ PREVENT	(1) to prevent, for preventing, helps prevent, in preventing, prevents, help prevent, prevent, prevented, can prevent, will prevent, preventing, at preventing, that prevents, may prevent (2) for relieving, relieves, relieve, to relieve, relieved, can relieve, in relieving, will relieve (3) to reduce, reduced, reduce, significantly reduces, significantly reduced, could reduce, reduces, can reduce (4) decreases, decreased, in decreasing, would decrease, diminishes, attenuates, lowers, cuts, eliminates, minimizes, to alleviate, alleviates, may ease (5) for treating, in treating, to treat, treated, to control, to combat, combats, in fighting, fighting, protects against (6) group had, group experienced, group reported (7) containing, based, improves, affects, helps, provides, provided makes, developed, remedy, was
SYNONYM Y	(1) see, refers to (2) also known as, aka, is known as (4) acute, mild, severe (5) also called, called, is also called (6) means, was defined as, ie
ISA (hypernym)	(1) is a new, a new (2) an, is an, is an effective (3) has, has an, (4) as, as an (5) see (6) exerts its (7) another, and other, or other, other, with other
ISA (hyponym)	(1) e.g, eg, e.g., (2) such as, like (3) including, include (4) i.e., ie, (5) activity, activity of (6) than, called (7) carbamezepine, drug, drugs, drugs such as, (8) effects of, effect of, properties of (9) efficacy of, action of, actions of, agents, agents such as (10) agent, agents, agents such as

The grouping shown in Table 2.1 was done manually by the authors to emphasize how groups of related patterns emerge from the automatic process.

Some future work is planned on automatically generating pattern variations and pattern synonyms.

Even if KPs were non-ambiguous (and most of them are not), they would not be failsafe access points to instances of the target semantic relation. For example:

- The KP's arguments do not represent domain-specific concepts and are not terminologically interesting.
- The KP's arguments are domain-specific, but it is not sure (experts could disagree) that the relation holds between its arguments.

These two kinds of noise are particularly challenging when extracting relation instances from uncategorized text, such as on the WWW. Since the communicative setting is unknown, it is not inconceivable that text fragments containing discourse between non-experts may "pollute" the data source. Non-experts are likely to use vague concepts like "problems" (e.g. "aspirin <may cause> severe problems") or instantiate relation instances which are simply incorrect (e.g. "1000mg of vitamin c, <aka> Ester C, if you feel a cold or flu coming on").

A key question addressed in this research is thus whether the advantages of using the entire web as a specialized corpus are not outweighed by disadvantages such as the retrieval of spurious instances like the above.

3. Description of the experiments

This section briefly presents the data used and describes the experiments.

3.1 Instances from UMLS

The UMLS knowledge sources comprise a Metathesaurus, a Semantic Network and a specialist lexicon all of which are continuously updated (the 2006AB edition is used in this paper) and made freely available by the US National Library of Medicine (http://umlsks.nlm.nih.gov). The Metathesaurus is a gigantic database containing information on more than 1 million biomedical and health related concepts and the semantic relations between them. The Semantic Network, which is an upper-level ontology used to synthesize and organize information in the Metathesaurus, currently contains no less than 54 relation types, including the 4 investigated in this paper.

Table 3.1 shows the target instances from UMLS used in our experiments with the number of terms and/or concepts included in the UMLS and some examples.

Table 3.1 – Target instances

Experiment	#Concepts for ? in UMLS	#Term variants for ? in UMLS	Examples
ISA(haloperidol,?)	6	42	antiemetic drugs dopamine antagonist major tranquilizers neuroleptic drug
ISA(?,antipsychotic)	82	205	aripiprazole clozapine haloperidol loxapine

INDUCES(?,vomiting)
INDUCES(?,emesis)
	38	90	citric acid
			ethanol
			ipecac syrup
			nitrous oxide

MAY_PREVENT(selenium,?)	1	1	deficiency diseases
SYNONYMY(glucose,?)	NA	3	dextrose d glucose d-glucose
SYNONYMY(formaldehyde,?)	NA	6	formalin methanal methyl aldehyde oxomethane
SYNONYMY(vitamin C,?)	NA	4	ascorbic acid c vitamin l ascorbic acid
SYNONYMY(progesterone,?)	NA	5	corpus luteum hormone luteal hormone pregnenedione

3.2 Corpora

The two corpora used are the WWW and BioMed. BioMed is an abbreviation for the BioMed Central's open access full-text corpus for text mining research. It contains some 24,000 articles of peer-reviewed biomedical research (http://www.biomedcentral.com/info/about/datamining) totalling 2.4 GBytes[1].

3.3 Method

For any semantic relation, we first establish a set of KPs using the automatic discovery approach of (Halskov and Barrière 2008) described in section 2. With such a set of KPs we can discover new relation instances by selecting an input term, leaving one argument blank and forming new queries based on the template, "<input term> <KP> <NP>". In the experiments discussed in this article the input terms are drugs, substances or symptoms from the UMLS Metathesaurus (see Table 3.1), KP belongs to P, P being the set of filtered patterns discovered for the target relation type (some examples previously shown in Table 2.1), and NP represents a sequence of NP chunk elements produced by tagging and chunking the term pair contexts using OpenNLP tools[2]. For each of the KPs in P the top 100 Yahoo contexts (text snippets) or all BioMed contexts (sentences) are returned.

4 Comparative experiments

First we look at the coverage of each corpus, and then we examine the results of applying the method (from section 3.3) to them.

4.1 Corpus coverage

Although UMLS is set as our gold standard, many of the terms present in UMLS for the experiments of interest in this paper, are not in BioMed. This makes their discovery as target instances of known relations impossible, but this is a problem caused by data sparseness in the corpus and not a shortcoming of the method used. For each experiment Table 4.1 lists: (column 1) the relation explored and the input term for that relation, (column 2) the number of term variants from UMLS in relation to the input term, (column 3) the number of term variants present in BioMed, and (column 4) the number of term variants present in the same article as the input term. As mentioned earlier, BioMed contains about 24,000 articles, and when finding relations between two terms, we hope to find them not only in the corpus as a whole but in the same article. Since a single occurrence may not provide enough

[1] These numbers were valid in March 2008 when the corpus was downloaded.

[2] http://opennlp.sourceforge.net/

evidence, column 5 in Table 4.1 shows how many UMLS term variants co-occur with the input term at least two times in the same article (file).

Table 4.1 – Presence of UMLS target instances in BioMed

Experiment	#Term variants for ? in UMLS	Present in BioMed	Present in same file (min 1 occ.)	Present in same file (min 2 occs)
ISA(haloperidol,?)	42	32	26	18
ISA(?,antipsychotic)	205	61	54	36
INDUCES(?,vomiting)	90	32	25	19
INDUCES(?,emesis)	90	32	13	12
MAY_PREVENT(selenium,?)	1	1	0	0
SYNONYMY(glucose,?)	3	3	3	3
SYNONYMY(formaldehyde,?)	6	2	2	2
SYNONYMY(vitamin C,?)	4	4	0	0
SYNONYMY(progesterone,?)	5	1	1	0

In comparison, the WWW coverage is shown in Table 4.2. While one or two occurrences in a specialized corpus like BioMed can be significant, the noisy and redundant nature of the WWW (identical pages being copied on multiple sites), as well as the large number of spurious pages returned by the search engine, leads us to experimentally establish a minimum of 10,000 hits when determining "presence". For each experiment Table 4.2 lists: (column 3) the number of term variants present on the WWW as established by the number of hit counts returned by a Yahoo query, and (column 4) the number of term variants present on the same web page as the input terms. The latter is established by using a joint query (input term AND term variant).

Table 4.2 – Presence of UMLS target instances on WWW

Experiment	#Term variants for ? in UMLS	Present in WWW > 10,000	In same page > 10,000
ISA(haloperidol,?)	42	41	36
ISA(?,antipsychotic)	205	88	49
INDUCES(?,vomiting)	90	66	53
INDUCES(?,emesis)	90	66	21
MAY_PREVENT(selenium,?)	1	1	1

Experiment	#Term variants for ? in UMLS	Present in WWW > 10,000	In same page > 10,000
SYNONYMY(glucose,?)	3	3	3
SYNONYMY(formaldehyde,?)	6	5	5
SYNONYMY(vitamin C,?)	4	4	4
SYNONYMY(progesterone,?)	5	4	2

Table 4.3 shows the comparative coverage of the WWW and BioMed as regards the occurrence of the UMLS term variants. Again, we only consider co-occurrences in same articles (or web pages). As shown, even with a relatively large minimum hit count of 10,000 pages for establishing WWW presence, the potential for discovery on the WWW is by far the highest. This was to be expected, but the quantitative analysis performed allows us to ground our intuition.

Table 4.3 – Comparative coverage of BioMed and WWW

Experiment	#Term variants for ? in UMLS	WWW coverage > 10000	BioMed coverage (min 1 occ)	BioMed coverage (min 2 occs)
ISA(haloperidol,?)	42	85.7%	61.9%	42.9%
ISA(?,antipsychotic)	205	23.9%	26.3%	17.6%
INDUCES(?,vomiting)	90	58.9%	27.8%	21.1%
INDUCES(?,emesis)	90	23.3%	14.4%	13.3%
MAY_PREVENT(selenium,?)	1	100%	0%	0%
SYNONYMY(glucose,?)	3	100%	100%	100%
SYNONYMY(formaldehyde,?)	6	83.3%	33.3%	33.3%
SYNONYMY(vitamin C,?)	4	100%	0%	0%
SYNONYMY(progesterone,?)	5	40.0%	20%	0%

4.2 Instance discovery

Table 4.4 presents the results of our instance discovery experiments on the BioMed corpus as follows: (column 1) the relation tested and the number of its KPs, (column 2) the input term (as the ISA relation can be found via KPs for hypernymy and hyponymy, results are split for haloperidol and antipsychotic), (column 3) the number of contexts found in BioMed containing the input term, (column 4) the number of contexts found in BioMed containing both the input term and a KP, (column 5) the number of NP candidates found in these contexts using the method described in section 3.3, and (column 6) the number of different candidates. We see that although the frequency (number of contexts) of some terms is quite high, their frequency in the presence of a KP is much lower. This is certainly indicative that BioMed has an expert-to-expert communicative setting and will not contain as many definitional contexts as an expert-to-novice setting would. The input term "glucose" is an illustrative example with over 10,000 occurrences in the corpus, but only 3 in the presence of a KP. This, of course, limits the number of NP candidates which can be found.

Table 4.4 – Number of contexts retrieved for each experiment in BioMed

Relation	Term	BioMed contexts with term	BioMed contexts with term + KP	Nb. candidate NPs	Nb. **different** candidates
ISA - 1 (hyper 16 KPs) - 2 (hypo 26 KPs)	Haloperidol-1	448	31	9	8
	Haloperidol-2	448	34	29	20
	Antipsychotic-1	1,491	117	79	59
	Antipsychotic-2	1,491	671	383	266
INDUCES (38 KPs)	vomiting	1,599	15	9	7
	emesis	225	23	2	2
MAY_PREVENT (72 KPs)	selenium	939	33	34	25
SYNONYMY (18 KPs)	glucose	11,227	3	5	5
	formaldehyde	2,525	2	1	1
	vitamin C	611	0	0	0
	progesterone	3,650	4	1	1

In a similar fashion, Table 4.5 lists results for the WWW. The number of text snippets found will differ depending on the number of KPs for the relation. For example, in the "NP <induces> vomiting" experiment a maximum of 3,800 snippets may be returned, as 38 KPs were discovered for this relation type. Although the query "term + KP" returns X snippets, not all of these actually contain the target strings. However, this problem is inherent to the way pages are indexed in the search engine and is thus beyond our control. In the "NP <induces> vomiting" example, only 2,259 of the 3,800 snippets returned actually contain a "KP + vomiting" in them. Also, the number in column 4 can be higher than the number in column 3 in cases where a term occurs in the same sentence with different overlapping patterns (e.g. "is a new antipsychotic" and "new antipsychotic"). As in Table 4.4 the total number of NP candidates is shown (column 5) and then the number of different candidates (column 6). The significant difference between the numbers in columns 5 and 6 shows how much redundancy there is on the WWW.

Table 4.5 – Number of snippets retrieved on WWW for each experiment

Relation	Term	Term present in snippet	Term + KP present in snippet	Nb. candidate NPs	Nb. **different** candidates
ISA - 1 (hyper 16 KPs) - 2 (hypo 26 KPs)	Haloperidol-1	586	177	173	73
	Haloperidol-2	1,976	1,015	964	270
	Antipsychotic-1	1,385	721	619	270
	Antipsychotic-2	1,930	2,000	1,096	596
INDUCES (38 KPs)	vomiting	3,108	2,259	1,983	453
	emesis	2,959	1,679	1,241	286
MAY_PREVENT (72 KPs)	selenium	4,918	823	850	334
SYNONYMY (18 KPs)	glucose	1,538	328	293	75
	formaldehyde	1,174	92	99	31
	vitamin C	1,025	71	69	26
	progesterone	1,358	161	164	57

Based on reduced gold standards determined by the respective coverage of the two corpora (Tables 4.1 and 4.2), Tables 4.6a and 4.6b show the precision/recall for each corpus. In order to measure recall/precision, we must count how many of the different candidates (last column of Tables 4.4 and 4.5) are valid, i.e. present in the gold standard. To consider a candidate valid, a relaxed matching algorithm is used (a candidate must contain a good UMLS term to be considered valid). This allows us to consider also more "expanded" candidates as valid, such as candidates containing non-essential adjectival modifiers (e.g. "haloperidol is a *classical* antipsychotic") and candidates which are part of conjunction, disjunction and/or ellipsis (e.g. "antipsychotics like haloperidol *or* chlorpromazine). Thus, Tables 4.6a and 4.6b list : (column 3) precision, (column 4) recall, (column 5) the actual number of term variants from UMLS covered by the corpus, (column 6) the “real recall” as if all of the term variants in the UMLS were to be found, and not just the subset of these terms actually present in each corpus.

Table 4.6a – Precision and Recall of WWW based on the UMLS as gold standard

Relation	Term	WWW precision	WWW recall	UMLS WWW coverage	"Real recall"
ISA - 1 (hyper 16 KPs) - 2 (hypo 26 KPs)	Haloperidol-1	8.2%	16.7%	36	14.3%
	Haloperidol-2	4.4%	33.3%	36	28.6%
	Antipsychotic-1	4.1%	22.4%	49	5.4%
	Antipsychotic-2	1.7%	20.4%	49	4.9%
INDUCES (38 KPs)	vomiting	0.4%	3.8%	53	2.2%
	emesis	0.7%	9.5%	21	2.2%
MAY_PREVENT (72 KPs)	selenium	0%	0%	1	0%
SYNONYMY (18 KPs)	glucose	1.3%	33.3%	3	33.3%
	formaldehyde	3.2%	16.7%	5	16.7%
	vitamin C	3.8%	25.0%	4	25.0%
	progesterone	0%	0%	2	0%

Table 4.6b – Precision and Recall of BioMed based on the UMLS as gold standard

Relation	Term	BioMed precision	BioMed recall	UMLS BioMed coverage	"Real recall"
ISA - 1 (hyper 16 KPs) - 2 (hypo 26 KPs)	Haloperidol-1	50.0%	15.4%	26	9.5%
	Haloperidol-2	10.0%	7.7%	26	4.8%
	Antipsychotic-1	8.5%	9.3%	54	2.4%
	Antipsychotic-2	3.4%	16.7%	54	4.4%
INDUCES (38 KPs)	vomiting	0%	0%	25	0%
	emesis	0%	0%	13	0%
MAY_PREVENT (72 KPs)	selenium	0%	0%	0	0%

Relation	Term	BioMed precision	BioMed recall	UMLS BioMed coverage	"Real recall"
SYNONYMY (18 KPs)	glucose	0%	0%	3	0%
	formaldehyde	0%	0%	2	0%
	vitamin C	NA	0%	0	0%
	progesterone	0%	0%	1	0%

A few examples of UMLS instances found on the WWW are: 1) the synonyms ascorbic acid - vitamin C, and dextrose - glucose, 2) ipecac syrup as inducing vomiting and emesis, 3) haloperidol as part of the family of neuroleptics and antipsychotic drugs, and 4) clozapine and trifluoperazine as examples of antipsychotics.

Overall, however, the results are a bit underwhelming. As regards the BioMed corpus, the candidates are rarely part of the UMLS leading to a recall of 0%. Although recall and precision numbers based on the WWW are also quite small, the recall here shows some potential. If recall is 0, as in most BioMed cases, there is nothing that can be gained from further analysis of the instance candidates found. But if recall is above 0, as with the WWW, methods for increasing precision can be envisaged.

5. Further WWW experiments

In this section we present two further experiments. First, we investigate a method to improve recall, which is simply to obtain a larger number of snippets for each experiment. Second, we suggest some simple candidate ranking methods to try to augment precision, but the challenge will be to do so without affecting recall. In Table 5.1, we show the impact of obtaining 250 snippets (columns 5 and 6) instead of 100 (columns 3 and 4 – copied from 4.6a).

Table 5.1 – Precision and recall with different numbers of snippets

Relation	Term	WWW 100 precision	WWW 100 recall	WWW 250 precision	WWW 250 recall
ISA - 1 (hyper 16 KPs) - 2 (hypo 26 KPs)	Haloperidol-1	8.2%	16.7%	5.3%	27.8%
	Haloperidol-2	4.4%	33.3%	2.8%	44.5%
	Antipsychotic-1	4.1%	22.4%	2.2%	24.7%
	Antipsychotic-2	1.7%	20.4%	1.2%	38.9%

Relation	Term	WWW 100 precision	WWW 100 recall	WWW 250 precision	WWW 250 recall
INDUCES (38 KPs)	vomiting	0.4%	3.8%	0.4%	7.5%
	emesis	0.7%	9.5%	0.5%	14.1%
MAY_PREVENT (72 KPs)	selenium	0%	0%	0%	0%
SYNONYMY (18 KPs)	glucose	1.3%	33.3%	0.6%	33.3%
	formaldehyde	3.2%	16.7%	2.0%	20.0%
	vitamin C	3.8%	25.0%	2.4%	25.0%
	progesterone	0%	0%	0%	0%

Recall is improved for the ISA relation, but not much for the other relations. As recall improves then precision, as expected, falls. When a small number of candidates are extracted, as for the BioMed corpus, we can easily imagine a user manually examining the limited number of instances found to establish their value, but with the WWW, a filtering system must be put in place, especially if we retrieve a larger number of snippets in the hope of finding more information. The challenge of such a system is to maintain its recall but have a better precision.

As a first investigation, a simple frequency ranking is used. We apply such filtering on the top 250 snippets. Results are shown in Table 5.2 when limiting the number of candidates to the top 50 and top 100 after ranking. We note how precision improves mostly for the ISA experiment, but recall diminishes, especially when we only look at the top 50.

Table 5.2 – Impact of frequency filtering on precision and recall

Relation	Term	ALL prec	100 prec	50 prec		All rec	100 rec	50 rec
ISA - 1 (hyper 16 KPs) - 2 (hypo 26 KPs)	Haloperidol-1	5.3%	9.0%	10%		27.8%	25.0%	13.9%
	Haloperidol-2	2.8%	10.0%	20%		44.5%	27.8%	27.8%
	Antipsychotic-1	2.2%	5%	8%		24.7%	10.0%	7.9%
	Antipsychotic-2	1.2%	8%	14%		38.9%	16.3%	14.2%
INDUCES (38 KPs)	vomiting	0.4%	1%	2%		7.5%	1.9%	1.9%
	emesis	0.5%	2%	2%		14.1%	9.4%	4.7%
MAY_PREVENT (72 KPs)	selenium	0%	0%	0%		0%	0%	0%

Relation	Term	ALL prec	100 prec	50 prec		All rec	100 rec	50 rec
SYNONYMY (18 KPs)	glucose	0.6%	1%	2%		33.3%	33.3%	33.3%
	formaldehyde	2.0%	1%	2%		20.0%	20.0%	20.0%
	vitamin C	2.4%	1%	2%		25.0%	25.0%	25.0%
	progesterone	0%	0%	0%		0%	0%	0%

As a second ranking method, we count the number of different KPs with which a candidate occurs. We revert to the previous ranking approach when a tie occurs. Table 5.3 shows comparative results of both ranking approaches with top 100 candidates: "freq" and "NbKP" being respectively the abbreviations for the first and second methods. The NbKP approach gives slightly better results, but more investigations should be performed to verify this.

Table 5.3 – Impact of Nb KPs filtering on Precision and Recall with top 100

Relation	Term	Freq precision	NbKP precision		Freq recall	NbKP recall
ISA - 1 (hyper 16 KPs) - 2 (hypo 26 KPs)	Haloperidol-1	9.0%	9.0%		25.0%	25.0%
	Haloperidol-2	10.0%	12.0%		27.8%	33.4%
	Antipsychotic-1	5%	7.0%		10.0%	14.2%
	Antipsychotic-2	8%	9.0%		16.3%	18.0%
INDUCES (38 KPs)	vomiting	1%	2.0%		1.9%	3.7%
	emesis	2%	2.0%		9.4%	9.4%
MAY_PREVENT (72 KPs)	selenium	0%	0%		0%	0%
SYNONYMY (18 KPs)	glucose	1%	1%		33.3%	33.3%
	formaldehyde	1%	1%		20.0%	20.0%
	vitamin C	1%	1%		25.0%	25.0%
	progesterone	0%	0%		0%	0%

5 Conclusion

We worked with 9 examples from the UMLS, covering 4 relation types (is-a, synonymy, may-prevent, induces). Our experiments attempted to rediscover the instances found for these 9 examples in the UMLS by searching in two different corpora: the WWW and BioMed. For the discovery process, we used a set of Knowledge Patterns extracted automatically on the WWW as implemented in the WWW2REL system (Halskov and Barrière 2008). Limitations of the BioMed materialized in terms of a poor coverage of the relation instances from the UMLS, and of knowledge patterns. This data sparseness problem is partially due to its expert-to-expert communicative setting.

Although the WWW is also limited in its coverage of UMLS instances, it showed a higher recall potential, and therefore we investigated further how to improve on the precision of our discovery approach while maintaining recall. Two simple filtering approaches were tested (frequency and nb. of KPs) to limit the list of candidates to more reliable ones. Depending on the application, a threshold could be set as a "reasonable" number of candidates to be looked at by a user. In our case we showed that a reduced set of ranked candidates (top 100) increased precision with a small loss of recall.

Overall, the results in terms of recall and precision versus the UMLS as a gold standard were poor and make us reflect on our evaluation exercise. The purpose of ontology expansion is knowledge discovery, and evaluating an ontology expansion methodology by its capacity to reproduce an existing ontology is only partially fair. Although we measured our results as a "rediscovery" exercise on the UMLS, it is difficult to show the full potential of a system in this manner. In the WWW2REL system, we did suggest an evaluation by human judges to rate system output. This would complement the "rediscovery" capability evaluated here. As human evaluation is quite costly, it should only be used in cases where the automatic rediscovery evaluation shows potential.

In conclusion, for ontology building, "going on the WWW" is important as recall is indeed boosted compared to a static repository (BioMed) which suffers from data sparseness problems. The challenge then becomes precision, and we presented a first encouraging path to be pursued in future research.

References

Agichtein, E. & L. Gravano; Snowball: Extracting relations from large plain-text collections, in *Proceedings of the Intl. Conference on Development and Learning (ICDL 2000)*. 2000.

Ahmad, K. & H. Fulford; *Semantic relations and their use in elaborating terminology*. Technical report, Uni. of Surrey, Computing Sciences. 1992.

Ahmad, K.; Pragmatics of specialist terms: The acquisition and representation of terminology, in *Machine Translation and the Lexicon, 3rd EAMT Workshop proceedings*. 1993.

Bowden, P. R. et al.; Extracting conceptual knowledge from text using explicit relation markers, in *Proceedings of the 9th European Knowledge Acquisition Workshop on Advances in Knowledge Acquisition*. 1996.

Gaizauskas, R. et al.; Protein structures and information extraction from biological texts: The PASTA system, *Bioinformatics*, 19(1), 2003.

Gillam, L. & M. Tariq & K. Ahmad; Terminology and the construction of ontology, *Terminology*, 11(1), 2005.

Girju, R. & D. Moldovan; Text mining for causal relations, in *Proceedings of the 15th Florida Artificial Intelligence Research Society conference*. 2002.

Halskov, Jakob & Caroline Barrière; Web-based extraction of semantic relation instances for terminology work. *Terminology*, 14(1). 2008

Hearst, Marti A; Automatic acquisition of hyponyms from large text corpora, in *Proceedings of COLING-92*. 1992.

Meyer, Ingrid; When terms move into our everyday lives: An overview of determinologization. *Terminology*, 6(1), 2000.

Meyer, Ingrid; Extracting knowledge-rich contexts for terminography, in D. Bourigault, C. Jacquemin, & Marie-Claude L'Homme (eds) *Recent Advances in Computational Terminology*, chapter 14, John Benjamins, 2001.

Mukherjea, S. & S. Sahay; Discovering biomedical relations utilizing the world-wide web, *Proceedings of Pacific Symposium on Bio-Computing*. 2006.

Pantel, P. & M. Pennacchiotti; Espresso: Leveraging generic patterns for automatically harvesting semantic relations, *Proceedings of ACL 2006*. 2006.

Yu, Hong & V. Hatzivassiloglou & C. Friedman & A. Rzhetsky & W. John Wilbur; Automatic extraction of gene and protein synonyms from MEDLINE and journal articles, in *Proceedings of the AMIA Symposium.* 2002.