Web-based extraction of semantic relation instances for terminology work

Halskov, Jakob; Barrière, Caroline

Publisher's version / Version de l'éditeur:
https://doi.org/10.1075/term.14.1.03hal
Terminology, 14, 1, pp. 20-44, 2008-08-01

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

Web-based extraction of semantic relation instances for terminology work

Jakob Halskov and Caroline Barrière

This article describes the implementation and evaluation of WWW2REL, a domain-independent and pattern-based knowledge discovery system which extracts semantic relation instances from text fragments on the WWW so as to assist terminologists updating or expanding existing ontologies. Unlike most comparable systems, WWW2REL is special in that it can be applied to any semantic relation type and operates directly on unannotated and uncategorized WWW text snippets rather than static repositories of academic papers from the target domain. The WWW is used for knowledge pattern (KP) discovery, KP filtering and relation instance discovery. The system is tested with the help of the biomedical UMLS Metathesaurus for four different relation types and is manually evaluated by four domain experts. This system evaluation shows how ranking relation instances by a measure of "knowledge pattern range" and applying two heuristics yields an average performance of 70% to 65% of the maximum possible F-score by top 10 and top 50 instances, respectively. Importantly, results show that much valuable information not present in the UMLS can be found through the proposed method. Finally, the article examines the domain-dependence of different aspects of the pattern-based knowledge discovery approach proposed.

Key words: automatic extraction of linguistic patterns, UMLS ontology expansion, web-based semantic relation extraction

1 Introduction

With the digital revolution and the genesis of a vast and freely accessible repository of text and knowledge known as the Internet, researchers from many fields, including text mining and computational terminology, are struggling to overcome a major challenge of the Internet Age, namely information overload. How does one find the gold nuggets of relevant knowledge washing down the information river?

In the context of computational terminology, especially for the task of

generating or updating ontologies and terminological knowledge bases, an important type of gold nuggets are semantic relations[1] expressed explicitly in natural language strings. Such strings have been called *knowledge-rich contexts* (KRCs). A KRC has been defined as:

> A context indicating at least one item of domain knowledge that could be useful for conceptual analysis. In other words, the context should indicate at least one conceptual characteristic, whether it be an attribute or a relation. (Meyer 2001: 281)

Hearst (1992) originally presented a way of identifying semantic relations in text by employing what has later been called *knowledge patterns* (Meyer 2001: 290), *knowledge probes* (Ahmad and Fulford 1992) or *explicit relation markers* (Bowden et al. 1996). In this article we will use the acronym *KP* for such patterns. For example, <to treat> in the following example is a KP which establishes a causal MAY_PREVENT relation between the two concepts represented by the terms *selenium* and *cancer*.

Revici was also an early advocate of using ***selenium*** *<to treat>* ***cancer****.*

KPs are used in pattern-based relation extraction systems which typically rely on techniques resembling the Dual Iterative Pattern Relation Expansion (DIPRE) algorithm introduced by Brin (1998) to extract, for example, instances of <company;location> tuples directly from the WWW. The basic idea of DIPRE-like approaches is shown in Figure 1.

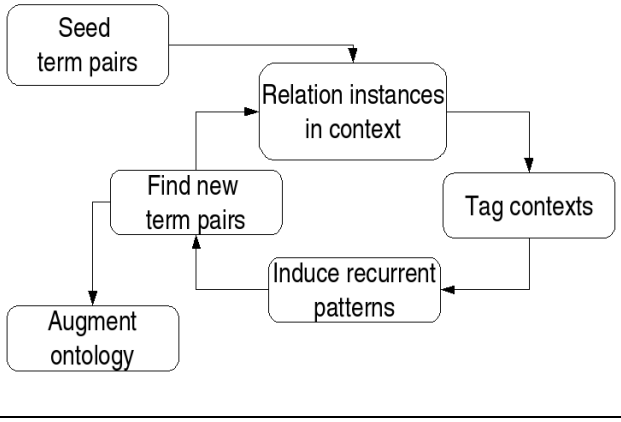

Figure 1: Dual Iterative Pattern Relation Expansion (DIPRE)

Figure 1. Dual Iterative Pattern Relation Expansion (DIPRE)

The intuition is that a small number of existing term pairs (seeds) instantiating the target semantic relation type can be used to induce patterns which occur in the term pair contexts, and these patterns can in turn be used to identify more term pairs and so forth. As Table 1 illustrates, useful KPs may occur in either the left, middle or right context of the term pair. Of course, KPs can also be discontinuous and span more than one of these contexts. Following the results of Agichtein and Gravano (2000), which show that the middle context is the most informative for English, we focus our first exploration on this context. While exploring the left and right contexts would be interesting, the fragmentary nature of the text snippets may render it difficult to extend the methodology in these directions.

Table 1. Knowledge patterns in context

Table 1: Knowledge patterns in context				
left	***term1***	*middle*	***Term2***	*right*
<causes of	*diarrhea*	**include>**	*parasites*	, some cancers ...
	diarrhea	**<induced by>**	*bacteria*	is a typical ...
to minimize the	*stomach irritation*		*Aspirin*	**<can cause>**
a **<side effect of**	*nicotinic acid*	**is>**	*flushing*	

As evidenced by Table 2, multiple pattern-based relation extraction systems have been developed since Hearst (1992). The column entitled *portability* in this table, which is by no means exhaustive, indicates the level of portability of the systems and the information in the parentheses explains what might cause this to be low (for example, the use of a biomedical Named Entity Recognition (NER) module). The column entitled *non-hierarch.* indicates whether the system handles non-hierarchical relation types, and *KP induction* indicates whether the patterns used by the system were introspectively devised or induced automatically from text.

Table 2. Pattern-based relation extraction systems

Table 2: Pattern-based relation extraction systems				
"system" (reference)	***portability***	***Non-hierarch.***	***KP induction***	***text source***
Espresso (Pantel and Pennacchiotti 2006)	high	Yes	yes	TREC-9
Snowball (Agichtein and Gravano 2000)	low (NER)	Yes	yes	newspapers
SGPE (Yu et al. 2002)	low (Biomedical. filters)	no	no	MEDLINE
PASTA (Gaizauskas et al. 2003)	Low (Biomedical lexicons)	yes	yes	MEDLINE
RelationAnnotator (Mukherjea and Sahay 2006)	low (Biomedical NER)	yes	no	WWW snip.
KnowItAll (Popescu et al. 2004, Etzioni et al. 2004)	high	no	no	WWW docs
(Nenadic and Ananiadou 2006)	low (Biomedical ontology)	yes	yes/no	MEDLINE
(Alfonseca et al. 2006)	low (NER)	yes	no	WWW docs
(Charniak and Berland 1999)	high	no	yes	newspapers
(Girju and Moldovan 2002)	Low (Wordnet)	yes	yes	TREC-9
WWW2REL	high	yes	yes	WWW snip.

Table 2 reveals that many systems already make use of text on the WWW (e.g., Etzioni et al. 2004; Popescu et al. 2004; Alfonseca et al. 2006; Mukherjea and Sahay 2006), but most systems are custom-tailored to a specific domain, for example biomedicine (e.g., Yu et al. 2002; Gaizauskas et al. 2003; Mukherjea and Sahay 2006; Nenadic and Ananiadou 2006), or focus on the Information Extraction IE task of extracting subclasses (e.g., Etzioni et al. 2004) or concept pairs which are "semantically related" via unlabelled relations (e.g., Nenadic and Ananiadou 2006). However, practical terminology work needs a wide range of labelled relations which may provide essential conceptual characteristics to be used in the writing of definitions.

This article thus describes the implementation and evaluation of WWW2REL, a relation extraction system which:

1. uses the WWW both for discovering KPs and relation instances;
2. can be applied to new domains;
3. can be applied to extract instances of any semantic relation type.

The combination of the three characteristics makes the system rather unique when compared to the systems listed in Table 2. It shares many features with the Espresso system, but uses the WWW rather than a controlled text source (TREC-9 competition). While section 2 discusses in more detail the properties of KPs, section 3 outlines the KP discovery and filtering step of WWW2REL. Section 4 describes the extraction of relation instances, and section 5 evaluates the performance of both individual ranking schemes but also of the system overall. Finally, portability issues are discussed in section 6, and section 7 summarizes our conclusions.

2 A closer look at KPs

KPs are not fail proof access points to instances of the target semantic relation. In a

landmark article, Meyer (2001) lists the following challenges to using KPs in automatic extraction tasks:

1. Unpredictability;
2. Polysemy;
3. Anaphoric reference;
4. Domain-dependence.

That KPs are unpredictable simply reflects the fact that they are part of natural rather than controlled or artificial language. There is virtually no limit to the creativity with which human beings express themselves, even when conveying specialized knowledge to each other. The polysemy, or ambiguity, of KPs is another fascinating feature of natural language (or annoying depending on the perspective). However, by using a reliability measure like KP range (to be described in section 4.1) it is possible to reduce the adverse effect of individual KP polysemy. Anaphoric reference is a third feature of natural language, notorious for its complexity and a hard nut to crack for natural language processing (NLP) applications. Resolving anaphora is a matter of boosting recall, and in the present implementation of WWW2REL this is not attempted because there is plenty of data to process (the entire WWW) and it makes sense to adopt a text mining approach. Finally, the possible domain-dependence of KPs will be addressed in section 6 of this article.

We suggest viewing the challenges of working with KPs in terms of the quality conditions these KPs should have, mainly:

1. High precision;
2. High recall;
3. High portability.

Striking the perfect balance between high recall, high precision and high portability can be hard, and also depends on the purpose of the application. However, to some extent the parameters are interdependent in that high precision KPs may tend to be domain-dependent and thus have a low portability and a low recall (at least in other domains than the one for which they were discovered). Conversely, highly portable KPs will tend to have a high recall and, presumably, a somewhat lower precision. Therefore the "discovery power" of individual KPs is likely to differ greatly.

For our application, WWW2REL extracts relation instances directly from the entire WWW and presents these to the user as ranked by their assessed reliability, so from a pragmatic viewpoint we favour precision and portability over recall even if we will later present a classical evaluation in terms of precision and recall.

In terms of precision-related challenges at least four noise-causing factors can occur when relying on KPs to extract relation instances automatically from text:

1. The KP does not instantiate a semantic relation at all.
2. The KP instantiates a different semantic relation than the target one.
3. The KP instantiates the target semantic relation, but its arguments do not represent domain-specific concepts and are not terminologically interesting.
4. The KP instantiates the target semantic relation, its arguments are domain-specific, but it is not sure that the relation holds between its arguments.

An example of the first cause of noise is the following sentence from a paper in a back-issue of *The Computer Journal*.[2]

What <is a> deep expert system?

The pattern <is a>, which might in other cases establish a hyponym-hypernym link, only establishes a link to the interrogative pronoun *what* and thus does not provide a

hypernym of the concept represented by *deep expert system*.

The second factor causing noise can be illustrated by the pattern <arise from> in the following sentence from a glossary of medical terms from the *New York Presbyterian Hospital*.[3]

> *Schwannomas and neurofibromas, tumors that <arise from> the sheaths that cover nerves and improve the conduction of nerve impulses.*

In general language *arise from* will almost always (except for poetic language, perhaps) be used to establish cause-effect relations, but in this case, however, it instantiates a locative relation between *tumors* and *sheaths*.

As for the third cause of noise, a more recent issue of *The Computer Journal*[4] provides another example.

> *We agree that this semantic decoupling <can cause> problems, although the degree of problems depend on the power of the weaving language and how it is used.*

This time the KP <can cause> does establish a causal link, but one argument (*problems*) is of such a general nature that the instance will presumably not be useful for terminologists who work bottom-up modelling the knowledge of special domains, or even sub-domains. This third cause of noise is perhaps the hardest to identify and evaluate, because the line between fuzzy categories and domain-specific concepts can be difficult to draw. Thus when extracting semantic relation instances for terminology an important subtask is automatic term recognition (ATR) which is typically achieved by a combination of linguistic and statistical techniques. WWW2REL includes a simple ATR technique to be described in section 4.2. (For more on ATR the reader may consult e.g., Drouin 2003.)

Finally, the fourth cause of noise is particularly relevant when using the WWW as a knowledge source. In neat collections of academic papers one would not expect to

encounter many incorrect semantic relations, but when authorship, text type and many other important quality parameters are unknown, it is not totally inconceivable that some relation instances will simply be false. The following is an example from a SYNONYMY experiment carried out as part of the WWW2REL evaluation.

1000mg of vitamin c, <aka> Ester C, if you feel a cold or flu coming on.

Since ester c is a modified (chemically enhanced) form of vitamin c, the SYNONYMY relation established by the KP <aka> is incorrect. The informal acronym for *also known as*, of course, signals that the communicative setting may not be an academic one, which could be one reason (but not a sufficient one) to have doubts about the authority of the source. This, of course, raises the question 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. Since the web is dynamic, multidisciplinary, multilingual, omnipresent and freely available, the authors find that the benefits outweigh the challenges. Admittedly, spurious relation instances may be more of a challenge in domains heavily affected by de-terminologization (Meyer and Mackintosh 2000).

Although we are aware of the different sources of noise, they will not be differentiated in our system analysis as the human evaluation in section 5 will not distinguish between them. Most of our filtering approaches will target noise reduction in general rather than particular types of noise. When a particular approach (such as the BNC heuristic discussed later) focuses on a particular type of noise, this will be highlighted. We postpone the discussion on the portability issue to section 6.

3 Discovering and filtering KPs

We now illustrate how KPs can be discovered using the WWW, but also how results

must be filtered to find "good" KPs given all the noise issues mentioned in section 2.

The first step is to select a small number of seed term pairs instantiating the target relation types. In the experiments four relation types are selected, two classical ones (ISA and SYNONYMY) and two relations important in biomedicine (MAY_PREVENT and INDUCES). All term pairs are randomly extracted from the Unified Medical Language System (UMLS) Metathesaurus which is freely available at http://umlsks.nlm.nih.gov (the 2006AB edition is used in the experiments). Figure 2 illustrates how seed term pairs for the ISA relation are extracted from a small fragment of the UMLS saving the concept *Antipsychotic* for the system evaluation.

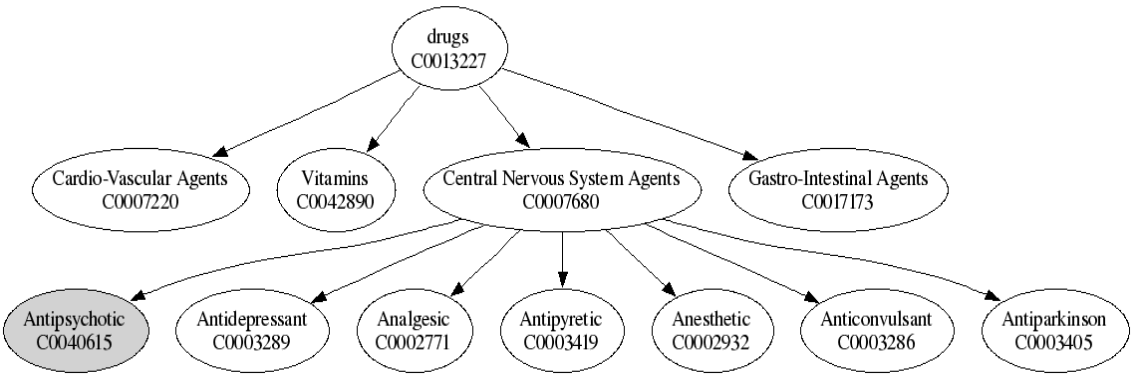

Figure 2: Using information in UMLS to discover KPs and augment UMLS

Figure 2. Using information in UMLS to discover KPs and augment UMLS

Google queries like the ones listed in Table 3 are then used to download the top 100 text snippets for each of the seed term pairs and thus compile four corpora of term pairs in context. The * is a word wildcard representing at least one word. As can be seen from the number of text snippets and tokens in each corpus, KP discovery is based on very small amounts of data. For SYNONYMY and the ISA relation additional query templates are used so as to capture also positional and morphological KP variants, e.g.

analgesics <such as> *ketamine*, but *ketamine* <is an> *analgesic*.

Table 3. Example queries and corpus sizes

Table 3: Example queries and corpus sizes					
Relation type	***example Google query***	***#seed pairs***	***#snippets***	***#tokens***	***#filtered KPs***
INDUCES	"carbon dioxide * headache"	40	4,054	94,000	71
MAY_PREVENT	"mineral oil * constipation"	40	3,993	91,000	101
SYNONYMY	"dyspnea * breathlessness" "breathlessness * dyspnea"	20 x 2	2,444	56,000	13
ISA	"ketamine * analgesic" "ketamine * analgesics" "analgesic * ketamine" "analgesics * ketamine"	40 x 4	9,519	205,000	41

For the two causal relations, inversing the term pairs would have led to the discovery of passive voice KPs. However, we decided to ignore those due to data sparseness. Although the WWW is vast, specialized biomedical term pairs co-occur relatively infrequently in themselves, and with the added restriction that they be connected with a passive verb phrase there was insufficient data to reliably discover KPs, especially for the MAY_PREVENT relation. For SYNONYMY and the ISA relation inversing argument order did not cause data sparseness.

The arbitrary number of 40 seed pairs for each relation was determined by the fact that for the INDUCES relation only 40 different UMLS term pairs co-occurred frequently enough on Google to meet the target of retrieving the top 100 text snippets for each pair. It is an interesting question whether a lower number of seed term pairs would reduce system performance at the stage of KP discovery and filtering. In most cases an equal amount of text snippets could presumably be retrieved simply by increasing the number of snippets downloaded for each term pair. However, the KP filtering technique called *term range* (to be described shortly) will presumably be more

effective the higher the number of seed pairs.

In order to reduce the set of KPs and eliminate noisy patterns induced from the Google text snippets, three strategies were developed:

1. Forcing a verb in the KP;

2. Assessing KP precision using negative term pairs;

3. Measuring the KP "term range."

The rationale for the first strategy is provided by studies like Barrière (2001) who empirically establishes that verbs are the most reliable form of KP. However, this first strategy might be too restrictive for certain relations, such as SYNONYMY and ISA. For example, it would eliminate perfectly valid KPs like <such as> and <or other>.

The second strategy is inspired from the Q-fold cross-validation technique used in machine learning. We split the positive pairs in 10 groups. We always look at the KPs found (learned) from 9 groups out of the 10 and test on the 10^{th} group. This process is repeated 10 times to average the results. Each time, for all the learned KPs, we measure their precision by counting how often they occur with the positive pairs (pairs expressing the desired relation) in the 10^{th} group versus a set of negative pairs (pairs expressing a different relation) of equal size. The Google hit counts for all $<t_1,KP,t_2>$ triplets (the numerators in the equation below) are normalized by taking into consideration how often the term pairs themselves co-occur on Google, i.e. $C_{Google}(t_1,*,t_2)$.

$$prec(KP) \approx \frac{\sum_{n=1}^{4} t_{1,n};t_{2,n} \in R_{pos} \frac{C_{Google}(t_{1,n},KP,t_{2,n})}{C_{Google}(t_{1,n},*,t_{2,n})}}{\sum_{n=1}^{4} t_{1,n};t_{2,n} \in R_{pos} \frac{C_{Google}(t_{1,n},KP,t_{2,n})}{C_{Google}(t_{1,n},*,t_{2,n})} + \sum_{n=1}^{4} t_{1,n};t_{2,n} \in R_{neg} \frac{C_{Google}(t_{1,n},KP,t_{2,n})}{C_{Google}(t_{1,n},*,t_{2,n})}}$$

As described above KP precision scores, prec(KP), are averaged over 10

iterations. While R_{pos} is a set of four positive term pairs for the target relation in a particular iteration, R_{neg} is a set of four negative term pairs in that same iteration. The negative pairs are selected at random from the UMLS Metathesaurus, but so that they represent semantically similar relation types and co-occur relatively frequently on the WWW (a minimum threshold of 15,000 hits was established empirically). An example of a negative pair for SYNONYMY is "cystic fibrosis;lung disease" which in fact represents an ISA relation.

The third strategy, KP term range, is in fact a by-product of the 10-fold cross-validation described above. It is to be understood as the number of iterations (i.e., groups of different term pairs) in which a particular KP candidate occurs. As much as the second strategy is meant to ensure that a KP is associated with a particular relation, the third strategy ensures that a KP is used often enough, with many different pairs to be relation-dependent and not pair-dependent.

The three filters are independent and can be activated or not depending on the semantic relation investigated. In our particular experiments, the verb filter was activated for the causal relations, and since this seemed highly effective, the term range filter was only activated for ISA and SYNONYMY. The precision filter was turned on for all semantic relations. Some thresholds were also set empirically and deserve further investigation in future work. The minimum average precision threshold was set to 50% and minimum term range to 4. The result of applying the different filters is four reduced sets of KPs, one for each relation studied. No attempt is made at generalizing the KPs, which is a popular way of boosting recall on small data sets. When using an unbounded and uncategorized text source, however, generalizing KPs might prove counter productive.

Table 4A. Example KPs for SYNONYMY, MAY_PREVENT and INDUCES

Table 4A: Example KPs with average precision > 50% and "term range" > 4 (SYNONYMY) or "forced verb" (the two causal relations)		
SYNONYMY	**MAY_PREVENT**	**INDUCES**
KP	**KP**	**KP**
or (10)	prevents	induces
see (9)	reduces	does not cause
also known as (9)	to prevent	can cause
ie (9)	prevent	to induce
means (8)	in preventing	Induced
also called (8)	had	Include
acute (7)	prevented	to cause
called (6)	decreases	causes
aka (6)	to treat	produces
is also called (5)	reduced	may cause
...	...	...

While Table 3 gives the total number of KPs (surviving the filtering) for each relation type, Table 4A gives actual examples of these KPs for the three relations of SYNONYMY, MAY_PREVENT and INDUCES.

Finally, Table 4B provides examples of the filtered ISA KPs as grouped by the four query variations or templates (see Table 3). While many of the ISA KPs in Table 4B may appear familiar and domain independent, quite a few seem specific to the biomedicine domain and would appear to have a low portability (e.g. "properties of", "efficacy of"). We discuss this further in section 6 with a small experiment which attempts to assess the portability of all KPs by applying them to another domain.

Table 4B. Example KPs for ISA

Table 4B: Example ISA KPs (by template) with average precision > 50% and "term range" > 4					
hypernym (plural) - hyponym	average precision	term range	***hypernym (singular) - hyponym***	average precision	term range
e.g.	100%	10	efficacy of	100%	9
such as	99.90%	10	action of	100%	9
including	99.20%	10	drugs	100%	9
like	89.60%	10	actions of	100%	8
i.e.	77.20%	9	agents	100%	7
include	69.00%	10	agents such as	100%	7
			called	100%	7
			drugs such as	100%	7

			properties of	100%	7
			effects of	100%	10
			...	...	...
hyponym-hypernym (singular)			***hyponym-hypernym (plural)***		
exerts its	100%	9	and other	99.01%	10
as an	100%	9	or other	69.48%	10
is an	100%	10	other	68.38%	10
is an effective	100%	10	with other	67.60%	10
an	99.96%	10	see	59.15%	10
has	89.57%	10	as	58.84%	10
is a new	79.83%	10			
is	76.72%	10			
a new	69.65%	10			
as	63.94%	10			

We do not perform a manual evaluation of the resulting KP lists, but rather proceed to an indirect evaluation by assessing their usefulness in the task of relation instance finding (what they were discovered for in the first place). We can see from Table 4B that some KPs (such as <is> or <is a>) are probably not too reliable in isolation, but if used in combination with other KPs (as we will see in the next section) they could prove useful. Certainly in an interactive system, we can imagine that a user could turn on/off the different KP filters (verb, precision, term range) and adjust the thresholds for each filter depending on the resulting lists obtained for diverse semantic relations and domains of interest. Also, future work will include a more in-depth assessment of the impact of changing the threshold values.

4 Extracting and ranking instances

For any semantic relation, after a set of KPs is established as described in the previous section, we can discover new relation instances on the WWW by selecting an input term, leaving one argument blank and forming new Google 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 (e.g. *aspirin*, *lactose*, *vomiting*), KP belongs to P, P being the set of filtered patterns discovered for the target relation type, and NP represents a sequence of NP chunk elements produced by tagging and chunking the Google snippets using *Treetagger*[5] and *Yamcha*.[6] For each of the KPs in P the top 100 snippets are returned. For example, in the "aspirin <INDUCES> X" experiment a maximum of 7,100 snippets may be returned as 71 KPs were discovered for this relation type.

4.1 Reliability measures

Four reliability measures are currently implemented in WWW2REL and are used to rank relation instances returned by the system. The measures are listed below.

1. frq(NP) = $\sum_{KP \in P} C_{sample}(t, KP, NP)$

2. kpr(NP) = $\left| \{ KP \in P \mid \exists (t, KP, NP) \} \right|$

3. fkpr(NP) = frq(NP)*kpr(NP)

4. pmi(NP) = $\dfrac{\sum_{KP \in P} \dfrac{pmi(t, KP, NP)}{\max_{pmi}} * r`(KP)}{|P|}$

The first measure, frq(NP), is a baseline ranking scheme in which the instances, or NPs, are simply ranked by their total co-occurrence frequency with the input term, t, and all KPs in the set of patterns, P, known by the system for the target relation type.

The second measure, kpr(NP), ranks NPs by the range of different KPs with which they occur in the sample. This will provide evidence that a suggested instance is expressed in different ways in the corpus. The third measure, fkpr(NP), is a hybrid of the two previous measures devised in an attempt at combining their strength.

Finally, the fourth measure, pmi(NP), is a slightly modified version of the

instance reliability measure, r(i), from the Espresso system (Pantel and Pennacchiotti 2006) in which a continuous KP reliability scale, r(KP), is used. In WWW2REL, r`(KP) is either 0 or 1 depending on whether the KP met the threshold values in the filtering stage. In the equation, pmi(t,KP,NP) is approximated by Google hit counts:

$$\frac{C_{Google}(t, KP, NP)}{C_{Google}(*, KP, *) * C_{Google}(t, *, NP)}$$

4.2 Heuristics

In addition to the four reliability measures, two heuristics are devised to help ranking the relation instances: BNC (British National Corpus) discounting and head grouping.

The BNC discounting heuristic is inspired by a statistical automatic term recognition technique commonly used in computational terminology, namely comparing the relative frequencies of a term candidate in a general language reference corpus versus the analysis corpus to compute a "weirdness" measure (Ahmad 1993). As "weirdness" is biased towards rare events, however, the heuristic in WWW2REL simply penalizes NPs whose head is too general and thus likely to be terminologically irrelevant. The BNC is used as model of general language, and the following formula illustrates how BNC filtering is applied to the "kpr" ranking scheme.

$$kpr_{bnc}(NP) = \frac{kpr(NP)}{\log(C_{BNC}(NP_{head}))}$$

The head noun grouping heuristic basically groups all instances by their NP head and executes a primary ranking of these heads based on the particular ranking scheme, e.g., "kpr." Using the same scheme all NPs sharing a particular head are then ranked. That grouping instances by their head noun could be a useful strategy for boosting performance will be illustrated by the example system output in section 5.3.

5 Evaluation

In this section, we attempt to measure the performance of WWW2REL at the task of knowledge discovery. The instance ranking approaches described in section 4 will be compared against a gold standard established by humans by using typical information retrieval measures, namely precision, recall and F-score (a combined measure). The section first discusses the evaluation setup and the issue of inter-annotator agreement. It then gives an example of system output and determines which relation instance ranking configuration yields the overall best performance. Finally, it examines to what extent the system can indeed augment the UMLS Metathesaurus.

5.1 Setup and purpose

As WWW2REL is essentially a knowledge discovery system which finds relation instances not recorded in the starting ontology, no existing gold standard can be used for evaluation. Although it is impossible to assess recall when looking for new knowledge on the WWW, a gold standard can be established by running the system for all the different experiments without any kind of instance ranking or filtering. Four experts are then asked to manually evaluate the correctness of each instance thus producing, themselves, a gold standard from unfiltered system output. All experts are young female graduates of the Danish University of Pharmaceutical Sciences (www.dfuni.dk). In total approximately 2,000 instances are assigned one of the following three judgments by each expert:

1. Relation instance is correct AND argument is domain-specific;
2. Unsure OR argument is fuzzy/vague;
3. Relation instance is incorrect.

In the gold standard, instances scoring an average judgment less than or equal to 1.50 are considered correct. For example, if three experts consider a relation instance

correct and one considers it incorrect, the system regards it as correct as (1+1+1+3)/4 is 1.50. When measuring system performance in terms of F-score, it is important to stress that recall is not to be understood in an absolute sense but only with respect to the number of relation instances in the samples retrieved by the system and receiving an average judgement of 1.50 or less.

Table 5. Experiments, recorded concepts in UMLS and correct instances in WWW2REL output

Table 5: Experiments, recorded concepts in UMLS and correct instances in WWW2REL output		
experiment	***#concepts for ? in UMLS***	***#correct/incorrect instances in system test***
ISA(haloperidol,?)	6	57/84
ISA(?,antipsychotic)	82	88/137
INDUCES(?,vomiting)	38	59/258
INDUCES(?,emesis)	38	25/51
INDUCES(aspirin,?)	0 (3)	148/217
MAY_PREVENT(selenium,?)	1	50/371
SYNONYMY(lactose,?)	NA	1/40
SYNONYMY(glucose,?)	NA	4/96
SYNONYMY(formaldehyde,?)	NA	4/42
SYNONYMY(vitamin C,?)	NA	5/58
SYNONYMY(progesterone,?)	NA	2/59

Table 5 lists the total of eleven different input terms used to test the performance of WWW2REL on the four different relation types. The numbers in the second column indicate how many concepts are already recorded in the UMLS Metathesaurus for each experiment. For the five SYNONYMY experiments *NA* indicates that this number is irrelevant when looking for synonyms of the same concept. The third column indicates how many positive and negative instances were found in the gold standards established by the experts. This number varies quite a lot among experiments as (a) the number of

KPs found for each semantic relation was different, (b) the number of instances found for each KP was different (with a maximum of 100) and (c) the number of instances annotated by the four judges varied as well.

As regards INDUCES(apirin,?) there are zero concepts recorded in the UMLS for ?. However, 3 concepts are recorded for the highly similar relation type HAS_PHYSIOLOGIC_EFFECT. Since we wish to investigate the INDUCES relation in its broader CAUSES sense, it would arguably be unfair not to consider these 3 concepts as well. Nevertheless, Table 5 suggests that a knowledge discovery process could indeed be useful to augment information in the UMLS Metathesaurus. This is especially the case for the two experiments involving aspirin and selenium.

Admittedly, the input terms represent a limited range of concepts, namely mainly drugs or substances with the exception of a single sign or symptom (i.e. vomiting/emesis). Future work should investigate input terms representing more diverse concepts.

5.2 Interannotator (dis)agreement

With four domain experts judging the correctness of the same relation instances, it is relevant to examine to what extent these experts are in agreement across all experiments. There are multiple measures of inter-annotator agreement, including joint probability, correlation coefficients and the kappa measure. Fleiss kappa measure (Fleiss 1971) is used in this case because it takes into account the amount of agreement expected to occur by chance and also works for more than just two annotators as opposed to Cohen's kappa (Cohen 1960). It is defined as $k = \frac{\overline{P} - \overline{P_e}}{1 - \overline{P_e}}$ where the numerator expresses the degree of agreement actually achieved and the denominator the maximum degree of agreement possible. A kappa value of 1 equals perfect agreement.

From the kappa values in Table 6 it is apparent that, overall, the four experts had a higher degree of agreement in experiments involving SYNONYMY and the ISA relation than in the case of the two causal relations. In particular, assessing the correctness of system proposals for positive effects of selenium seemed to cause the experts considerable trouble. Private talks with the experts revealed that a range of possibly beneficial physiological effects of selenium intake have been proposed, but that there is scientific agreement on the correctness of only a very few of these effects. Also, applying WWW2REL to find entities which induce vomiting resulted in the lowest inter-annotator agreement of all experiments. This is due to the much greater search space of the WWW which causes the results to contain many instances of the target relation which are correct but whose arguments are judged to be too general and thus irrelevant to the domain (e.g., *too much chocolate* <can cause> *vomiting*). As indicated by the kappa values in the Table 6, however, exchanging *vomiting* by the more technical (domain-specific) synonym *emesis* doubled inter-annotator agreement.

Table 6. Inter-annotator agreement across all experiments

Table 6: Inter-annotator agreement across all experiments			
No.	***Experiment***	***kappa***	***observations***
1	SYNONYMY(formaldehyde,?)	0.75	46
2	ISA(?,antipsychotic)	0.62	225
3	ISA(haloperidol,?)	0.57	141
4	SYNONYMY(lactose,?)	0.57	41
5	SYNONYMY(glucose,?)	0.56	100
6	SYNONYMY(vitamin C,?)	0.45	63
7	INDUCES(?,emesis)	0.42	76
8	SYNONYMY(progesterone,?)	0.40	61
9	INDUCES(aspirin,?)	0.28	365
10	MAY_PREVENT(selenium,?)	0.28	421
11	INDUCES(?,vomiting)	0.23	317

5.3 Example output

Table 7. "Aspirin <INDUCES> X" - top 10 candidates

Table 7: "Aspirin <INDUCES> X" – top 10 candidates						
Rank	***candidate ("frq")***	***judges***	avg.	***candidate ("fkpr-bnc-head")***	***judges***	avg.
1	apoptosis	1,1,2,1	1.25	bleeding	1,1,2,2	1.5
2	bleeding	1,1,2,2	1.5	gastrointestinal bleeding	1,1,1,1	1.0
3	asthma	1,1,2,1	1.25	stomach bleeding	1,1,1,1	1.0
4	ulcers	1,2,2,1	1.5	more bleeding	1,1,2,3	1.75
5	ringing	1,2,2,3	2.0	internal bleeding	1,1,2,1	1.25
...	...	...	...	...	...	...

Table 7 provides an example of system output given the input term *aspirin* and the set of KPs discovered for the INDUCES relation. On the left hand side, candidate instances are ranked by simple frequency in the corpus of Google text snippets (baseline "frq"), but on the right hand side ranking is done by the "fkpr" scheme with the two heuristics turned on (described in section 4). The main advantage of the "fkpr-bnc-head" ranking appears to be that its head grouping helps to cluster information on synonyms and hyponyms, which is useful to terminologists. Depending on the application, we could also view the diversity offered by the top ranking candidates on the left of the table ("frq") as quite informative to the terminologist although somewhat noisier. If equipped with a simple graphical user interface, WWW2REL could make it possible for terminologists to switch back and forth between viewing only candidate heads and viewing all candidates sharing a particular noun head.

5.4 Performance of ranking schemes

As illustrated by Figure 3, simply ranking relation instances by their frequency is indeed a suboptimal, baseline strategy. The main weakness of the "frq" scheme is that it quickly trails off in terms of precision compared to the "kpr" scheme, for example.

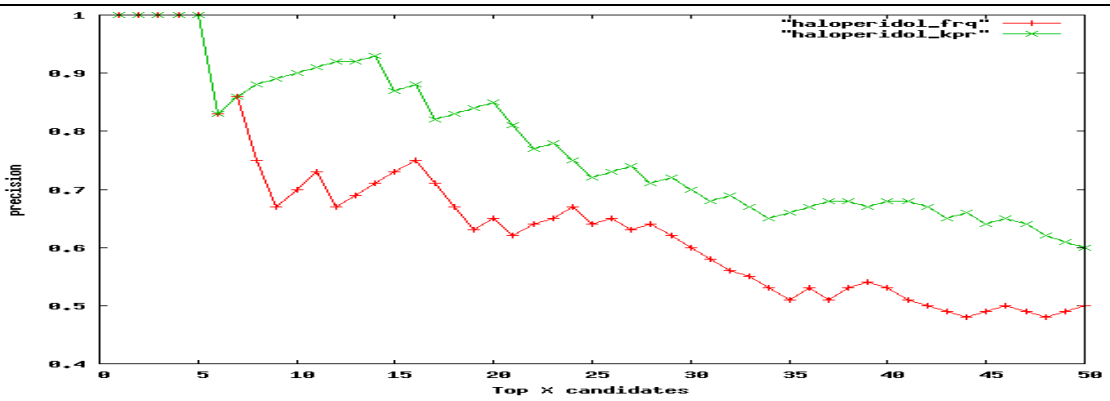

Figure 3: Precision of "frq" versus "kpr" - "haloperidol ISA X"

Figure 3. Precision of "frq" versus "kpr" - "haloperidol ISA X"

Presented with a wide range of relation instance ranking schemes applied to a number of different experiments, the intended users of the system are likely to request a default setting which will produce the overall best results. The average performance of the individual schemes across all experiments and all relation types can be computed as the proportion between actual F-score and the maximum possible F-score at a uniform cut-off point of e.g., the top 10 and top 50 instances, respectively. In the "aspirin INDUCES X" experiment the maximum possible F-score by the top 50 instances is given by:

$$F_{\max.pos.}(50) = 2 * \frac{\frac{50}{50} * \frac{50}{148}}{\frac{50}{50} + \frac{50}{148}} \approx 0.51$$

as there are a total of 148 correct instances in this experiment.

Table 8. Overall best ranking scheme (across all experiments)

Table 8: Overall best ranking scheme (across all experiments)			
top 10 instances		***top 50 instances***	
$\frac{F(10)}{F\max.pos.(10)}$	***scheme***	$\frac{F(50)}{F\max.pos.(50)}$	***scheme***

Table 8: Overall best ranking scheme (across all experiments)			
top 10 instances		***top 50 instances***	
$\frac{F(10)}{F\max.pos.(10)}$	***scheme***	$\frac{F(50)}{F\max.pos.(50)}$	***scheme***
70%	kpr-bnc-head	66%	fkpr-bnc-head
69%	kpr-head	65%	kpr-bnc-head
65%	frq-bnc-head	61%	frq-bnc-head
64%	fkpr-bnc-head	59%	kpr-bnc
59%	kpr-bnc	57%	kpr-head

We can now measure the average performance of each possible ranking configuration across all eleven experiments by computing $\frac{\sum_{e=1}^{n} \frac{F(i)}{F_{\max.pos.}(i)}}{n}$ where e is a counter representing the individual experiments, n=11, and the number of instances, i, is either 10 or 50. From the average performance scores in Table 8 it is apparent that the best configurations are those in which both heuristics, i.e. head grouping ("head") and BNC discounting ("bnc"), are activated. It also appears that ranking instances by their KP range results in the best overall performance (cf. "kpr-bnc-head"), even if the fkpr scheme is slightly better when considering the top 50 instances.

5.5 Augmenting the UMLS

This section examines more closely the usefulness of WWW2REL towards augmenting an existing terminological resource, namely the UMLS Metathesaurus upon which the system was trained and tested in the previous sections. It is no trivial matter to automatically measure the proportion of "new" or unrecorded, knowledge retrieved. While instances which differ only in terms of non-essential adjectival modifiers (e.g., *bleeding* vs. *severe bleeding*) should arguably not be counted as pieces of new knowledge, the novelty of synonyms (e.g., *gastro-intestinal bleeding* vs. s*tomach*

bleeding) is perhaps another matter. Moreover, technical issues like ensuring a reliable automatic decomposition of NPs containing conjunctions or disjunctions may disturb the picture. For these reasons the proportion of new knowledge retrieved by WWW2REL was **manually** computed based on the following guidelines.

1. Novelty by means of non-essential adjectival modifiers is ignored;
2. Novelty through conjunction, disjunction and/or ellipsis is ignored;
3. Novelty by means of synonymy is accepted.

Table 9. New knowledge retrieved by WWW2REL

Table 9: New knowledge retrieved by WWW2REL		
experiment	***% of correct instances not in UMLS***	***examples of unrecorded, correct instances***
MAY_PREVENT(selenium,?)	100%	lung cancer risk, DNA damage ...
INDUCES(aspirin,?)	100%	ulcers, stomach irritation ...
INDUCES(?,vomiting)	95%	meningitis, overeating ...
INDUCES(?,emesis)	96%	apomorphine, carboplatin, cisplatin...
ISA(haloperidol,?)	69%	CNS-depressant drug ...
ISA(?,antipsychotic)	39%	amisulpride, bretazenil, fluoxetine ...
SYNONYMY(lactose,?)	100%	milk sugar
SYNONYMY(glucose,?)	33%	corn sugar
SYNONYMY(formaldehyde,?)	20%	metylene oxide
SYNONYMY(progesterone,?)	100%	progestin, progestogen
SYNONYMY(vitamin c,?)	33%	ascorbate

Observing these three principles, Table 9 lists the proportion of all correct instances returned by WWW2REL for each experiment but not recorded in the UMLS.

Looking at the percentages in Table 9 (and the number of positives in Table 5) we see that new instances are primarily retrieved for the causal relation types. Indeed, nearly 100% of all causal instances judged to be correct by the experts are not recorded in the UMLS. For the ISA relation the proportion of new knowledge is somewhat lower, especially the coverage of antipsychotic drugs in the UMLS appears fairly comprehensive (only 39% of all correct instances are not recorded). As regards the

relatively high proportion of unrecorded hypernyms proposed for *haloperidol* (69%), this can be explained by the fact that quite a few of these hypernyms (e.g., *medication* or *medicine*) are so superordinate that they are on the point of being terminologically irrelevant and thus do not really augment the UMLS.

With the two exceptions of *progesterone* and *lactose*, the proportion of new knowledge is considerably lower in the SYNONYMY experiments. One reason is that the number of synonyms for a particular concept is much lower, and presumably also more fixed, than e.g. the number of new drugs of a particular category or the number of side effects discovered for a particular drug.

The results in Table 9 speak for themselves and show how knowledge not included in the UMLS can be semi-automatically discovered on the WWW based on seed term pairs originating from the UMLS.

6 KP portability

All the methods presented for KP discovery, KP filtering and relation instance discovery are not dependent on the biomedical domain, although that domain was chosen as a privileged testing domain given the availability of the UMLS Metathesaurus which allowed thorough system tests. Although the portability of WWW2REL as a knowledge discovery methodology is really more important than the portability of individual KPs, we now examine the latter as it is interesting to see to what extent KPs need to be rediscovered for each new domain. Assessing a pattern's true portability would require an extensive analysis of its use in multifarious domains, but even if this section only features an analysis of a single other domain, this should at least give some indication of the domain dependence of each KP as well as elucidate more global portability issues.

The domain used in this comparison is information technology (IT). IT was selected for two reasons. Firstly, it provides a good contrast to biomedicine, because the terminology of IT has a much higher density of terms with homographs in non-target domains. IT terms are often formed by semantic extension of existing general language lexical items, for example *window*, *icon* and *desktop*. Secondly, in the absence of a panel of IT experts the authors are able to act as surrogate experts for this domain.

Table 10. Applying WWW2REL to another domain (IT)

Table 10: Applying WWW2REL to another domain (IT)		
Experiments	***Query flexibility***	***#snippets***
ISA(perl,?)	morphological + positional variation	3,513
ISA(?, programming language)	morphological + positional variation	2,393
MAY_PREVENT(firewall,?)	morphological variation	7,131
INDUCES(computer virus,?)	morphological variation	3,353
SYNONYMY(subroutine,?)	positional variation	1,005

Five small experiments, listed in Table 10, were carried out to test the domain specificity of the filtered KPs for the four relation types. This section will summarize the findings, but also provide a single example of WWW2REL output which illustrates how not just KPs but also ranking heuristics may have low portability.

Table 11. "perl ISA X" - Activating BNC discounting in the IT domain

Table 11: "perl ISA X" - Activating BNC discounting in the IT domain			
candidate ("kpr-head")	***judgment***	***candidate ("kpr-bnc-head")***	***judgment***
language	2	oopl	1
programming language	1	tools o'reilly	3
scripting language	1	improbables	3
interpreted language	1	envt ajaxtk	3
object-oriented language	1	esoterica	3
interpreted programming language	1	lanuage	2
oo language	1	optimizer	3
implementation language	1	envt	3
network-capable high-level language	1	tools amazonuk	3
...	...	...	...

As revealed by the two sets of top hypernym candidates of *perl* in Table 11,

activating the BNC-based heuristic has an absolutely disastrous effect on precision and is not portable from the domain of biomedicine to IT. The problem is that the NP head, *language* gets heavily penalized because it is very frequent in the BNC, albeit in the sense "natural language" rather than "artificial language."

However, when turning BNC discounting off, the other heuristic of head grouping proves a useful mechanism not just for boosting precision, but also because it provides not just one, but several different hypernyms (e.g., *object-oriented language*) and their synonyms (e.g., *oo language*). Just by studying the top few candidates as ranked by "kpr-head" a terminologist can glean much useful information for building an ontology of programming languages. Indeed, other systems have analyzed the semantics of modifier-head relations in NPs to accomplish exactly this (e.g., Gillam et al. 2005), and future versions of WWW2REL should include a module for identifying implicit ISA relations.

Table 12 lists the proportion of KPs used in the eleven biomedical experiments which also make a contribution in the five corresponding IT experiments. When comparing the percentages for the different relation types, it appears that SYNONYMY KPs are the most domain-independent ones of the four sets learned for biomedicine. ISA KPs also appear to be relatively portable, while KPs for the two causal relations, especially the INDUCES relation, are the most domain-dependent of all. For example, verbs like *relieve*, *ease*, *cure*, *treat*, *provoke* and *produce* do not appear to be useful markers of causality in IT, although they are highly efficient in biomedicine.

Table 12. Domain-dependence of KPs discovered for biomedicine

Table 12: Domain-dependence of KPs discovered for biomedicine	
Relation	***% of biomedical KPs also used in IT experiments***
SYNONYMY	91.7% (11/12)
ISA	82.9% (34/41)
MAY_PREVENT	68.5% (50/73)

INDUCES	32.9% (23/70)

7 Conclusion

In conclusion, the evaluation of WWW2REL showed that it is possible to discover KPs and extract both hierarchical and non-hierarchical semantic relation instances from semantically unannotated and uncategorized text fragments on the WWW with high precision. Specifically, using an instance reliability measure of KP range (kpr) proved efficient, and applying the two heuristics of head grouping and BNC discounting further boosted the overall performance. The UMLS Metathesaurus augmentation experiments indicated that instances of the causal relation type are harder to extract with high precision than ISA, but it must be mentioned that the average precision in the causal experiments is heavily reduced by the "selenium MAY_PREVENT X" experiment which had an extremely low inter-annotator agreement.

An important purpose of knowledge discovery systems is the expansion of existing resources. We clearly saw how the UMLS could, in fact, be augmented. Relatively little information on causal relations involving *aspirin* or *selenium* was recorded in the UMLS, and WWW2REL was able to find several such instances.

Our comparative experiment between biomedicine and IT led to interesting findings. (a) although the methodology for KP discovery does not depend on the relation type, it is very important to take term pairs from the domain of study because some KPs will tend to be domain specific, (b) it is important to understand the nature of the semantic relations of interest to decide on the use of certain filters for discrediting KP candidates (e.g. forcing a verb), (c) it is important to understand the nature of the target domain, as BNC discounting, for example, may be counter productive in domains where terms are formed through semantic extension of existing general language

lexemes. Terminologists who are aware of these problems may fine tune the various parameters of WWW2REL and even deviate from the default setting to optimize their results.

Notes

[1] Semantic relations are understood as a hypernym of conceptual relations. Synonymy is an example of a semantic relation type which is not also a conceptual relation.
[2] *The Computer Journal* 40(4), Oxford University Press, 1997
[3] http://www.nyp.org/health/spinal-tumors.html
[4] *The Computer Journal* 46(5), Oxford University Press, 2003
[5] www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[6] http://chasen.org/~taku/software/yamcha/

References

Agichtein, E. and L. Gravano. 2000. "Snowball: Extracting relations from large plain-text collections." In *Proceedings of the fifth ACM conference on Digital libraries*. 85-94. San Antonio, Texas.

Ahmad, K. and H. Fulford. 1992. *Semantic Relations and their Use in Elaborating Terminology*. Technical report, University of Surrey, Computing Sciences.

Ahmad, K. 1993. "Pragmatics of specialist terms: The acquisition and representation of terminology." In *Machine Translation and the Lexicon, 3rd EAMT Workshop Proceedings*. 51-76. Heidelberg, Germany.

Alfonseca, E., P. Castells, M. Okumura and M. Ruiz-Casado. 2006. "A rote extractor with edit distance-based generalization and multi-corpora precision calculation." In *Proceedings of ACL 2006*. 9-16. Sydney, Australia.

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

Bowden, P.R., P. Halstead and T.G. Rose. 1996. "Extracting conceptual knowledge from text using explicit relation markers." In *Proceedings of the 9th European Knowledge Acquisition Workshop on Advances in Knowledge Acquisition*. 147-162. Nottingham, UK.

Brin, S. 1998. "Extracting patterns and relations from the world wide web." In *Proceedings of the International Conference on Extending Database Technology (EDBT 1998)*. 172-183. Valencia, Spain.

Charniak, E. and M. Berland. 1999. "Finding parts in very large corpora." In *Proceedings of ACL 1998*. 57-64. Montreal, Canada.

Cohen, J. 1960. "A coefficient for agreement for nominal scales." In *Education and Psychological Measurement* 20: 37-46.

Drouin, P. 2003. "Term extraction using non-technical corpora as a point of leverage."

Terminology 9(1): 99-117.

Etzioni, O., M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Sodeland, D. Weld and A. Yates. 2004. "Web-scale information extraction in KnowItAll." In *Proceedings of the 13th international conference on WWW*. 100-110. New York.

Fleiss, J. L. 1971. "Measuring nominal scale agreement among many raters." In *Psychological Bulletin* 76(5): 378-382.

Gaizauskas, R., G. Demetriou, P.J. Artymiuk and P. Willett. 2003. "Protein structures and information extraction from biological texts: The PASTA system." *Bioinformatics* 19(1): 135-143.

Gillam, L., M. Tariq and K. Ahmad 2005. "Terminology and the construction of ontology." *Terminology* 11(1): 55-81.

Girju, R. and D. Moldovan. 2002. "Text mining for causal relations." In *Proceedings of the 15th Florida Artificial Intelligence Research Society conference*. 360-364. Pensacola, Florida.

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

Meyer, I. and K. Mackintosh. 2000. "When terms move into our everyday lives: An overview of de-terminologization." *Terminology* 6(1): 111-138.

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

Mukherjea, S. and S. Sahay. 2006. "Discovering biomedical relations utilizing the world-wide web." In *Proceedings of Pacific Symposium on Bio-Computing*. 164-175. Maui, Hawaii.

Nenadic, G. and S. Ananiadou. 2006. "Mining semantically related terms from biomedical literature." *ACM Transactions on Asian Language Information Processing* 5(1): 22-43.

Pantel, P. and M. Pennacchiotti. 2006. "Espresso: Leveraging generic patterns for automatically harvesting semantic relations." In *Proceedings of ACL 2006*. 113-120. Sydney, Australia.

Popescu, A.-M., A. Yates and O. Etzioni. 2004. "Class extraction from the world wide web." In *Proceedings of the AAAI 2004 Workshop on Adaptive Text Extraction and Mining*. San Jose, California.

Yu, H., V. Hatzivassiloglou, C. Friedman, A. Rzhetsky and W.J. Wilbur. 2002. "Automatic extraction of gene and protein synonyms from MEDLINE and journal articles." In *Proceedings of the AMIA Symposium 2002*. 919-923. San Antonio, Texas.