Discovering Cross-Ontology Subsumption Relationships by Using Ontological Annotations on Biomedical Literature

Watson W.K. Chua[1]*and Jung-jae Kim[1]

[1]Nanyang Technological University
School of Computer Engineering
Block N4 #02a-32, Nanyang Avenue
Singapore 639798

ABSTRACT

Cross-ontology concept subsumption relationships facilitate the integration of ontologies by explicitly defining the generalization of a concept over other concepts in different ontologies. However, existing methods for discovering these relationships show poor performances and one of the problems is the lack of instance data in ontologies that can be used to identify cross-ontology subsumptions reliably. To address the problem, we present a novel method, *SURD* (SUbsumption Relation Discovery), which uses annotations on biomedical text corpora for populating ontologies with instances. Subsumption relationships between pairs of concepts are then determined based on their shared instances. *SURD* shows good performance when applied to biomedical ontologies, achieving precision values of 0.786 and 0.729 for cross-ontology subsumptions between the ontology pairs *GRO-UMLS Metathesaurus* and *GENIA-UMLS Metathesaurus* respectively. As a practical application, we used *SURD's* subsumptions for automated ontological corpus annotation and achieved F-measures of 0.693 and 0.783 on the GRO and GENIA corpora respectively. These results are superior to the results of using subsumption relations inferred from equivalence relations (F-measures of 0.569 and 0.645) and subsumption relations identified with Hearst patterns (F-measures of 0.002 and 0.096).

1 INTRODUCTION

Ontologies are continually being developed to model sub-domains of the biomedical sciences. While the proliferation of ontologies brings about greater expressivity in knowledge representation, it also creates a new problem in knowledge sharing. Applications using different ontologies face inter-operability issues when relationships (e.g. equivalence, subsumption) between concepts in different ontologies are not explicitly stated. The process of discovering relationships between concepts in different ontologies is known as Ontology Alignment (Euzenat and Shvaiko, 2007) and extensive research has been performed in this area. However, the majority of existing Ontology Alignment research focuses solely on the discovery of equivalence relationships between concepts in different ontologies. Equivalence relationships alone are insufficient to fully support inter-operability, especially in the Biomedical Science domain where ontologies are often orthogonal (Ghazvinian *et al.*, 2010) and few concepts from different ontologies are semantically equivalent. Subsumption relationships complement the equivalence relationships by explicitly stating the generalization of a concept over other concepts. In this paper, our objective is to find subsumption relationships directly, without inferring them from equivalence relationships.

For example, consider the integration of two populated ontologies, O_1 and O_2, and a query of listing all instances of C_i in the integrated ontology where C_i is originally a concept in O_1. The answers of the query consist of the instances of C_i and its descendants in O_1, as well as the instances of O_2 concepts which are semantic equivalences and descendants of C_i. We can find some of the O_2 instances by using equivalence relations between O_1 and O_2 concepts. For example, if there is an ancestor concept of C_j, C_{ja}, which is equivalent to C_i or one of its descendant concepts, C_{id}, we can infer through description logics that $C_j \sqsubseteq C_i$:

$$C_j \sqsubseteq C_{ja} \equiv C_{id} \sqsubseteq C_i$$

However, it is not always possible to find equivalent concepts C_{id} and C_{ja} that act as bridges for inferring meaningful subsumption relations. For example, if two ontologies are orthogonal, it is most likely that only a few top-level concepts (e.g. *Thing*) are shared by the two ontologies, and the subsumption relations of such generic concepts are not very useful for knowledge sharing. We will show that there are many subsumption relations between biomedical ontologies that cannot be deduced from equivalence relations and that they are highly useful for inter-operability between ontologies.

We propose an instance-based technique, *SURD* (short for SUbsumption Relations Discovery) to find cross-ontology subsumption relations directly. This technique determines whether a subsumption relationship exists between a pair of concepts based on the common instances they share. However, many existing ontologies are schema ontologies with no instances (Ehrig *et al.*, 2005). We resolve this issue using a novel technique that uses ontological annotations on biomedical literature as instances. Apart from ontology integration, subsumption relations can also be used for (semi-) automatic ontology annotation. This is most useful in the biomedical domain in which comprehensive lexical resources like UMLS exist. We further elaborate this application in Section 5.

The outline of the paper is as follows: Related work is presented in Section 2. Section 3 describes the *SURD* technique that finds subsumption relations between different ontologies while in Section 4, we describe the experiments we carried out and present the results. We show applications of our technique in Section 5, before finally concluding the paper in Section 6.

*To whom correspondence should be addressed: watsonchua@pmail.ntu.edu.sg

2 RELATED WORK

The discovery of relationships between concepts in different ontologies has been extensively researched and previous works have been surveyed in (Euzenat and Shvaiko, 2007) and (Chua and Goh, 2010). Techniques like *ASMOV* (Jean-Mary *et al.*, 2009) use logical inference on equivalence relationships to find subsumption relationships as explained in the Introduction. The subsumption relations found using these techniques have poor coverage as there are many cases where subsumption relationships can exist without equivalence relationships. We address this issue by proposing an instance-based approach. van Hage *et al.*, 2005 make use of text corpora for finding subsumption relations by using *Hearst* patterns (Hearst, 1992). While *Hearst* patterns work well in the general domain, their effectiveness is limited when the domain is restricted to a specialized domain like the biomedical domain where authors can safely assume that readers have sufficient background knowledge. Therefore, the type of an entity is seldom explicitly specified using such *Hearst* patterns as "C_i such as/including/especially C_j" where C_i and C_j are concepts from different ontologies. We tested the *Hearst* patterns on 1200 biomedical documents and could find less than 20 subsumption relations between concepts from ontologies of our interest (see Section 5 for details). In contrast, *SURD* shows much better coverage for biomedical ontologies.

Instance-based methods (Doan *et al.*, 2004; Kirsten *et al.*, 2007) have been used for finding equivalence relationships between ontologies. However, these techniques cannot be widely applied due to the difficulties in finding common sets of instances shared by ontology-pairs. By using different ontological annotations on the same set of biomedical documents, we are able to alleviate the problem and allow instance-based techniques to be applied for finding both equivalence and subsumption relationships.

Spiliopoulos *et al.*, 2010 use machine learning to find subsumption relations in the absence of instances. A model is trained using intra-ontology subsumption relations in a pair of ontologies before applying the trained model to concept-pairs in the two ontologies to find cross-ontology subsumption relations. The approach, known as Classification-Based Learning of Subsumption Relations (CSR), is effective if both ontologies have similar hierarchical structures. However, the biomedical ontologies we analyze in this paper have quite different structures: For example, the UMLS Metathesaurus' hierarchical structure is rather flat, considering its size, but GRO has a relatively deep hierarchical structure. The *SURD* approach does not face this problem since it is able to populate the ontologies with instances and is not heavily dependent on the structures of ontologies.

3 METHODOLOGY

Given a pair of ontologies O_i and O_j, we want to find all triplets $< C_i, R, C_j >$ where $C_i \in O_i$, $C_j \in O_j$ and $R \in \{\equiv, \sqsubset, \sqsupset, \perp\}$. $C_i \equiv C_j$ means that C_i and C_j are equivalent concepts. $C_i \sqsubset C_j$ indicates that C_i is a sub concept of C_j, while $C_i \sqsupset C_j$ means the inverse subsumption relation. $C_i \perp C_j$ means C_i and C_j have no subsumption or equivalence relationship. *SURD* discovers equivalence and subsumption relationships by populating the ontologies with instances from textual annotations and using heuristics based on the shared instances of two concepts to determine if a subsumption relationship exists between them. We give an outline of *SURD* in Figure 1.

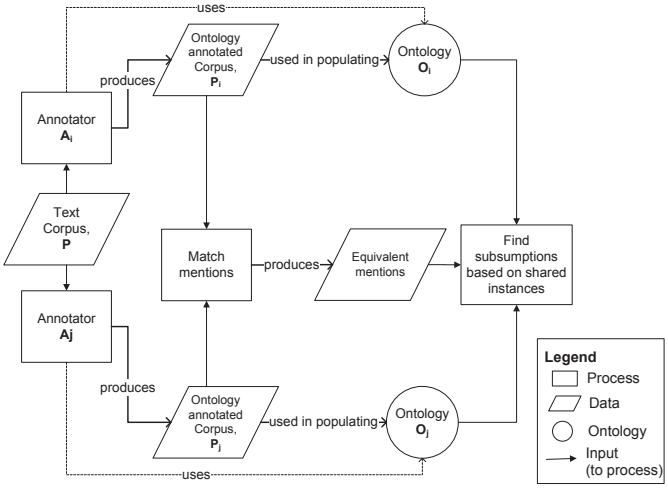

Fig. 1. Overview of the Subsumption Relations Discovery (*SURD*) technique

3.1 Using Annotations as Instances

We propose to populate ontologies with annotated phrases of text from publicly available biomedical literature. Two types of resources can be used: 1) Corpora manually annotated with ontology concepts (e.g. GENIA Corpus (Kim *et al.*, 2003), GRO Corpus (Kim *et al.*, 2011)) and 2) automatic ontology concept annotations by using tools such as MetaMap (Aronson, 2001), which is based on the UMLS Metathesaurus, and the NCBO Annotator (Jonquet *et al.*, 2009), an ontology-independent tool. In this paper, we compare the GENIA and GRO corpora, which are based on the GENIA ontology and the Gene Regulation Ontology (GRO) respectively, with the UMLS Metathesaurus-based annotations by MetaMap.

Given a pair of ontologies O_i and O_j, a corpus P consisting of n documents, and two annotators A_i and A_j which annotate the documents using O_i and O_j, respectively, we annotate the documents in P using A_i and A_j to get two annotated corpora P_i and P_j. Each phrase (or mention) m_α in document d_β of P annotated with a concept $C_\gamma \in O_{i(j)}$ is then made an instance of C_γ. Essentially, our objective is to compare two sets of annotations on the same set of documents to find out which pairs of concepts are frequently used to annotate the same mentions as subsumption relationships potentially exist between these pairs.

The first step of our method is to locate mentions annotated by both annotators. This is not always straightforward because different annotators have different guidelines for annotation, particularly with regards to mention boundary. For example, given the sentence "*spi-B, like spi-1, was found to be expressed in various murine and human hematopoietic cell lines...*", annotator A_1 might annotate the mention *hematopoietic cell lines* with the concept *Cell* while annotator A_2 might annotate the mention *cell lines* with the concept *Cell Line*. In this example, *cell lines* and *hematopoietic cell lines* refer to the same semantic entity and thus, the concepts *Cell* and

Cell Line share the same mention and form a candidate pair for subsumption. We consider two mentions to be matches if they have the same head noun.

3.2 Finding Subsumptions based on Shared Instances

We use an indicator known as the Co-Annotation Ratio (CAR) to determine subsumption relationships. The Co-Annotation Ratio between two concepts $C_p \in O_p$ and $C_q \in O_q$, CAR_{pq}, computes the ratio of mentions annotated with both C_p and C_q to the number of mentions annotated with C_p, and is expressed by the following equation:

$$CAR_{pq} = \frac{|\{m_p|m_p : C_p, m_q : C_q, m_p \equiv m_q\}|}{|\{m_p|m_p : C_p\}|}$$

where the notation $m : C$, defines m to be an instance of concept C and $m_p \equiv m_q$ is true if m_p and m_q are matched in the mention-matching step. CAR is asymmetric and we compare CAR_{ij} and CAR_{ji} in order to determine the relationship between C_i and C_j. We propose the following heuristics using the two indicators to determine the concept relation:

1. If CAR_{ij} is high and CAR_{ji} is low, then $C_i \sqsubset C_j$
2. If CAR_{ij} is low and CAR_{ji} is high, then $C_i \sqsupset C_j$
3. If CAR_{ij} is high and CAR_{ji} is high, then $C_i \equiv C_j$
4. If CAR_{ij} is low and CAR_{ji} is low, then $C_i \perp C_j$

The first rule states that, if a large proportion of the instances (or mentions) belonging to C_i also belongs to C_j but only a small proportion of the instances belonging to C_j belong to C_i, then C_i is highly possible to be a subsumee of C_j. In fact, this is a relaxation to the definition of a subsumption relation which depicts that $C_\alpha \sqsubset C_\beta$ if and only if all instances of C_α are also instances of C_β. We introduce the relaxed heuristic in order to enhance the sensitivity of subsumption relation identification. The second rule represents the converse of the first. The third rule states that if a large proportion of the instances belonging to C_i also belong to C_j and vice versa, we consider the two concepts to be equivalent. This method of statistical analysis allows us to identify both equivalence and subsumption relations at the same time. Lastly, if both C_i and C_j have large proportions of instances which are not common, C_i and C_j are most likely disjoint. In *SURD*, we consider a CAR ≥ 0.5 to be *high* and a CAR < 0.5 to be *low*.

4 EXPERIMENT AND RESULTS

Experiments were performed on two pairs of ontologies using two different corpora. The first is the GRO corpus with 200 *PubMed* abstracts and the second is the 2008 version of the GENIA corpus consisting of 1000 *PubMed* abstracts. The two corpora were manually annotated by human experts with concepts of biological entities from two ontologies, namely the Gene Regulation Ontology (Beisswanger *et al.*, 2008), and the GENIA ontology (Kim *et al.*, 2006), respectively. 10400 mentions in the GRO Corpus were manually annotated using 212 unique GRO concepts while 54533 mentions were manually annotated using 46 unique GENIA concepts. We then annotated each corpus automatically using *MetaMap* to get annotations based on the UMLS Metathesaurus. 17119 mentions in the GRO Corpus were annotated using 3063 unique UMLS concepts by *MetaMap* while 99626 mentions in the GENIA Corpus were annotated using 5796 unique UMLS concepts.

Since the Metathesaurus is not a formal ontology[1], we adopted the *OWL* version of the Semantic Network ontology[2] and extended it by adding the Metathesaurus concepts used by *MetaMap* in the annotation of the GRO Corpus and the GENIA Corpus to get two ontologies, $UMLS_{GRO}$ and $UMLS_{GENIA}$, respectively[3]. *SURD* is then used to discover subsumptions between GRO and $UMLS_{GRO}$, and also between GENIA and $UMLS_{GENIA}$. Henceforth, we use *UMLS* to indicate either $UMLS_{GRO}$ or $UMLS_{GENIA}$ when the corpus being referred to is clear.

The three ontologies were chosen because they have different granularities. The GENIA ontology is a coarse-grained ontology with leaf concepts which are general like *Protein Molecule* and *Carbohydrate*. On the other hand, the Metathesaurus has a wide coverage and its leaf concepts are mostly fine-grained and very specific (e.g. *p56* and *Glucose*). Therefore, we can expect to find many subsumption relations between the two ontologies. GRO is relatively coarse-grained, as compared to the Metathesaurus, but has very specific concepts regarding the domain of gene regulation. We show that *SURD* works well with the domain-specific ontology GRO as well as with the generic GENIA ontology.

The output of *SURD* is a set of triplets $A_{SURD}(O_i, O_j) = \{< C_i, R, C_j >\}$. We evaluated this output $A_{SURD}(O_i, O_j)$ by measuring the precision of a randomly chosen subset $A^R_{SURD}(O_i, O_j)$ through manual validation by a biologist. We were not able to measure the recall as we do not have a complete reference set of subsumption relations between *GRO-UMLS* and *GENIA-UMLS*. Since subsumption relations can be inferred from equivalence relations, we also mark the triplets in $A_{SURD}(O_i, O_j)$ which are inferrable from equivalence relations. This allows us to find the discovered subsumption relations which are not redundant and truly useful. The steps taken in our evaluation are as follows:

1. Finding relations in $A_{SURD}(O_i, O_j)$ that are inferrable from equivalence relations
   a. We use *BOAT* (Chua and Kim, 2012), a matcher for finding equivalence relations, to find all equivalence relations between O_i and O_j to get $A_{BOAT}(O_i, O_j)$.
   b. $A_{BOAT}(O_i, O_j)$ is expanded to include subsumption relations which can be inferred from equivalence relations. For each equivalence correspondence $C_i \equiv C_j$, we add the relations $C_{id} \sqsubset C_{ja}$ and $C_{ia} \sqsupset C_{jd}$ for all C_{id} and C_{jd}, descendants of C_i and C_j, respectively, as well as all C_{ia} and C_{ja}, which are ancestors of C_i and C_j, respectively.
   c. Triplets in $A_{SURD}(O_i, O_j)$ that are found in $A_{BOAT}(O_i, O_j)$ are marked as inferrable.
2. Estimating the precision of $A_{SURD}(O_i, O_j)$
   a. Randomly select a subset of n triplets $A^R_{SURD}(O_i, O_j)$, from $A_{SURD}(O_i, O_j)$.
   b. A biologist familiar with the domain and ontologies involved examines each concept pair (C_i, C_j) from triplet $< C_i, R, C_j > \in A^R_{SURD}(O_i, O_j)$ and assigns a relation $R_m \in \{\sqsubset, \sqsupset, \equiv, \perp\}$ to the pair.

[1] http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

[2] http://krono.act.uji.es/people/Ernesto/UMLS_SN_OWL

[3] The extended ontologies are available at `http://nlp.sce.ntu.edu.sg/SURD`

c. A triplet $< C_i, R, C_j >\in A^R_{SURD}(O_i, O_j)$ is correct if $R = R_m$.

d. Precision = $\frac{\text{Number of correct triplets in } A^R_{SURD}(O_i, O_j)}{n}$

Table 1 shows examples of trivial (inferrable from equivalence correspondences) and non-trivial subsumption correspondences found by *SURD* where the source concept (from GRO or GENIA) subsumes the *UMLS* concept in the same row. The correspondence in row 1 is trivial because *Cell* from *GRO* is equivalent to *Cell* in the Semantic Network (SN). Since *Blood Cell* is a subtype of Cell, it can be inferred that *Blood Cell* $\sqsubseteq$ *SN:Cell* $\equiv$ *GRO:Cell* $\rightarrow$ *Blood Cell* $\sqsubseteq$ *GRO:Cell*. On the other hand, it is not possible to infer the subsumption correspondence in row 3 because *NF-kappa B* is not a subclass of any concept that matches *GRO:TranscriptionFactor* or any of its descendants. *SURD*'s strength is that it is able to discover subsumption correspondences even if there are no equivalence correspondences linking concepts in the two ontologies. The list of correspondences found using *SURD* can be downloaded from our project Web site at `http://nlp.sce.ntu.edu.sg/SURD`.

Table 2 shows our evaluation results where we randomly selected subsets of extracted relations and asked the biologist to manually validate them due to the lack of reference dataset. *SURD* shows high precisions of 0.786 and 0.729 for finding subsumption relations between *GRO-UMLS* and *GENIA-UMLS* respectively. Note that significant proportions of the relations correctly identified (42.3% for *GRO-UMLS* and 86.6% for *GENIA-UMLS*) are non-trivial. The percentage of correct non-trivial relations for *GENIA-UMLS* is higher than that for *GRO-UMLS* because much fewer equivalence relationships were discovered between *GENIA* and *UMLS* (19) as compared to those for *GRO* and *UMLS* (79). With the high precision and non-triviality on both ontology pairs, we used *SURD* for automatic corpus annotations which we describe in the next section.

No.	Source Concept	Source Ont	UMLS Concept	Trivial
1	Cell	GRO	Blood cell	Y
2	Virus	GENIA	Sendai Virus	Y
3	TranscriptionFactor	GRO	NF-kappa B	N
4	Peptide	GENIA	Enkephalin	N

Table 1. Examples of trivial and non-trivial subsumption relations found by *SURD*

	Relations Found	Relations Validated	Correct	Prec.	Correct non-trivial
GRO-UMLS	1952	514 (26%)	404	0.786	171 (42.3 %)
GENIA-UMLS	5200	790 (15 %)	576	0.729	499 (86.6 %)

Table 2. Results of validation on relations found by *SURD*

5 AUTOMATIC CORPUS ANNOTATION

We make use of the subsumption and equivalence relations discovered by *SURD* for automatic ontological corpus annotation. This application is dependent upon *UMLS*, which is a well-known lexical resource in the biomedical domain. By using *MetaMap*, which automatically recognizes *UMLS* terms in text, we can locate the instances of the *UMLS Metathesaurus* concepts and then link them to the corresponding concepts of GENIA ontology and GRO through the subsumption and equivalence relations. Note that we can infer generalizations, but not specificities. In other words, we can use $C_{GRO} \sqsupseteq C_{UMLS}$ relations, but not $C_{UMLS} \sqsupset C_{GRO}$ relations. For example, we can infer that a mention *NF kappa B* annotated with the concept *Transcription Factor* is also a *Protein* if *Transcription Factor* $\sqsubset$ *Protein* is true. However, we cannot infer anything about the same mention from the subsumption *SOX9* $\sqsubset$ *Transcription Factor*. The automatic annotation of text with GRO concepts is a three-step process, as shown in Figure 2. Sentences are first automatically annotated with *UMLS Metathesaurus* concepts using *MetaMap*. Next, *GRO* concepts which are equivalent to or more general than the UMLS concepts are retrieved from the correspondences found by *SURD*. Lastly, the retrieved GRO concepts are used to annotate the mentions which their corresponding UMLS concepts were used to annotate.

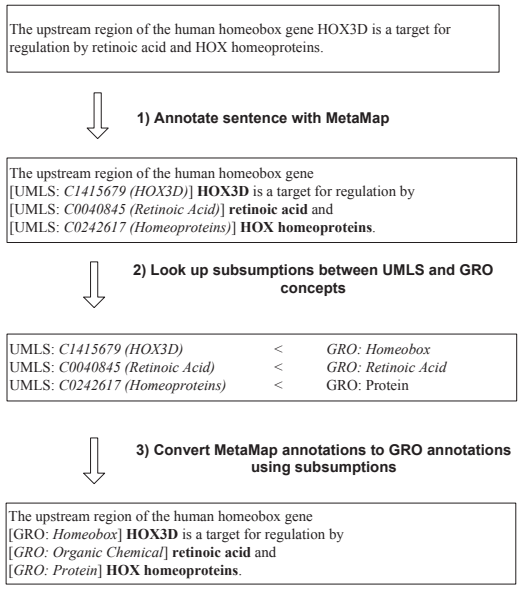

Fig. 2. Example of automatic ontological corpus annotation

To evaluate the performance, we use a cross-validation approach where Precision, Recall and F-Measure are used as performance measures. The following steps were carried out to evaluate *SURD* when used for the automatic annotation of GRO Corpus with GRO concepts by using *MetaMap*. The process was repeated for the annotation of GENIA corpus with GENIA ontology concepts by using *MetaMap* and any reference to *GRO* can be replaced with *GENIA* for the second experiment.

1. Split the document sets, P_{UMLS} and P_{GRO} (i.e. the GRO corpus annotated with UMLS and GRO concepts, respectively), into 10 folds: $F_0, F_1, \ldots F_9$.
2. For each iteration t from 0 to 9, find the set of triplets $A^t_{SURD}(GRO, UMLS)$ using *SURD*, based on the annotations in the documents in all folds but F_t.

3. Filter $A^t_{SURD}(GRO, UMLS)$ to keep only those triplets $< C_{GRO}, R, C_{UMLS} >$ where $R \in \{\equiv, \sqsupset\}$
4. For each mention m_i of fold F_t annotated with C_{UMLSi} by *MetaMap* in P_{UMLS}, we annotate m_i with each concept $C_{GROi\alpha}$ where $< C_{GROi\alpha}, \sqsupseteq, C_{UMLSi} > \in A^t_{SURD}(GRO, UMLS)$.
5. Consequently, a mention m_i in fold F_t of P_{UMLS} is annotated with zero or more GRO concepts, which forms a set $X_i = \{C_{GROi\alpha} | C_{UMLSi} \sqsubseteq C_{GRO_i\alpha}\}$. We use the reference concept C'_{GROi} to determine if m_i is cross-annotated correctly, where C'_{GROi} is the concept used to manually annotate m'_i, the matching mention of m_i in P_{GRO}. We find the average Precision, Recall and F-Measure over all folds for a corpus containing k mentions using:
   - Precision = $\frac{\Sigma_i^k |\{C_{GROi\alpha} \in X_i | C_{GROi\alpha} \sqsupseteq C'_{GROi}\}|}{\Sigma_i^k |X_i|}$
   - Recall = $\frac{\Sigma_i^k |\{C'_{GROi} | \exists C_{GROi\alpha} \in X_i, C_{GROi\alpha} \sqsupseteq C'_{GROi}\}|}{\Sigma_i^k |\{C'_{GROi}\}|}$
   - F-Measure = $\frac{2 \times Precision \times Recall}{Precision + Recall}$

Automatic annotation was also performed using the subsumption relations found by *BOAT* (as described in 4) as well as those found using *Hearst* patterns. We were not able to compare with CSR as the tool is not publicly available. After manually validating all the *Hearst* patterns in the two corpora, we were able to find only 3 subsumption relations for *GRO-UMLS* and 14 for *GENIA-UMLS*.

Table 3 shows the performance of automatic ontology annotation using subsumption relations from the three techniques. Though the annotations made using *BOAT*'s inferred relations have the highest recall for the GRO Corpus, it is achieved at the cost of precision. Similarly, the high precision of the *Hearst* method on the GENIA Corpus is achieved at the expense of recall. *SURD* has the highest precision for the GRO Corpus, the highest recall for the GENIA Corpus, and the highest F-measures for both corpora. The last row of Table 3 shows the average distance between reference annotation C'_{GRO} and its closest match C_{GRO} identified using automatic annotation, for all C'_{GRO}s that have matches. An average distance close to 0 means that the automatic annotations are mostly identical to the manual annotations, while a large average distance means that mentions are often matched to more general concepts than those of manual annotations. *SURD* has very small average distances for both corpora. This shows that in addition to having higher coverage, automatic annotations using *SURD* are almost identical to the manual annotations.

	GRO_UMLS			**GENIA_UMLS**		
	SURD	*BOAT*	*Hearst*	*SURD*	*BOAT*	*Hearst*
Precision	0.866	0.491	0.8	0.839	0.843	0.896
Recall	0.577	0.664	0.001	0.735	0.539	0.050
F-Measure	0.693	0.565	0.002	0.783	0.658	0.096
Avg. Dist	0.038	1.4	0	0	3.1	1.0

Table 3. Automatic ontological corpus annotation results

6 CONCLUSION AND FUTURE WORK

We have presented a novel technique for discovering cross-ontology subsumption relations which uses ontological annotations on biomedical text corpora to determine subsumption relations between concepts that share mentions. The relations discovered are highly precise and have wide coverage and can thus be used for integrating a pair of ontologies with minimal expert curation. We also showed that they can be effectively used for automated cross-ontology annotations on biomedical corpora. For future work, we plan to apply *SURD* to other biomedical ontologies and also complement the equivalence relations found by *BOAT* with the subsumption relations found by *SURD* for the integration of ontologies populated with annotations so as to effectively perform semantic querying on biomedical literature.

ACKNOWLEDGEMENTS

The authors would like to thank Xu Han for helping to validate the subsumption relations.

REFERENCES

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap Program. In *Proceedings of the AMIA Annual Symposium (AMIA 2001)*, pages 17–21, Portland, Oregon, USA.

Beisswanger, E., Lee, V., Kim, J.-J., Rebholz-Schuhmann, D., Splendiani, A., Dameron, O., Schulz, S., and Hahn, U. (2008). Gene Regulation Ontology (GRO): Design principles and use cases. In *Studies in Health Technology and Informatics*, volume 136, pages 9–14.

Chua, W. W. K. and Goh, A. E. S. (2010). Techniques for discovering correspondences between ontologies. *International Journal of Web and Grid Services*, **6**(3), 213–243.

Chua, W. W. K. and Kim, J.-J. (2012). BOAT: Automatic alignment of biomedical ontologies using term informativeness and candidate selection. *Journal of Biomedical Informatics*, **45**(2), 337–349.

Doan, A., Madhavan, J., Domingos, P., and Halevy, A. (2004). Ontology matching: A machine learning approach. In *Handbook on Ontologies in Information Systems*, pages 397–416. Springer.

Ehrig, M., Staab, S., and Sure, Y. (2005). Bootstrapping ontology alignment methods with APFEL. In *Proceedings of the 4th International Semantic Web Conference (ISWC 2005)*, pages 186–200, Galway, Ireland. Springer.

Euzenat, J. and Shvaiko, P. (2007). Ontology matching. Springer-Verlag.

Ghazvinian, A., Noy, N. F., and Musen, M. A. (2010). How orthogonal are the OBO Foundry Ontologies? In *Proceedings of Bio-Ontologies 2010: Semantic Applications in Life Sciences*, pages 164–167, Boston, USA.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, pages 539–545, Nantes, France.

Jean-Mary, Y. R., Shironoshita, E. P., and Kabuka, M. R. (2009). Ontology matching with semantic verification. *Journal of Web Semantics*, **7**(3), 235–251.

Jonquet, C., Shah, N. H., and Musen, M. A. (2009). The Open Biomedical Annotator. In *Proceedings of the AMIA Summit on Translational Bioinformatics*, pages 56–60, San Francisco, CA, USA,.

Kim, J.-D., Ohta, T., Tateisi, Y., and ichi Tsujii, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**(suppl 1), i180–i182.

Kim, J.-D., Ohta, T., Teteisi, Y., and Tsujii, J. (2006). GENIA ontology (TR-NLP-UT-2006-2). Technical report, Tsujii Laboratory, University of Tokyo.

Kim, J.-J., Han, X., and Chua, W. W. K. (2011). Annotation of biomedical text with Gene Regulation Ontology: Towards semantic web for biomedical literature. In *The Fourth International Symposium on Languages in Biology and Medicine (LBM 2011)*, pages 63–70.

Kirsten, T., Thor, A., and Rahm, E. (2007). Instance-based matching of large life science ontologies. In *Proceedings of Data Integration in the Life Sciences (DILS)*, pages 172–187.

Spiliopoulos, V., Vouros, G. A., and Karkaletsis, V. (2010). On the discovery of subsumption relations for the alignment of ontologies. *Journal of Web Semantics*, **8**(1), 69–88.

van Hage, W. R., Katrenko, S., and Schreiber, G. (2005). A method to combine linguistic ontology-mapping techniques. In *Proceeedings of 4th International Semantic Web Conference (ISWC 2005)*, pages 732–744, Galway, Ireland.