Semi-Supervised Learning to Identify UMLS Semantic Relations

Yuan Luo¹, Ozlem Uzuner¹²
¹Massachusetts Institute of Technology ²State University of New York at Albany

Abstract
The UMLS Semantic Network is constructed by experts and requires periodic expert review to update. We propose and implement a semi-supervised approach for automatically identifying UMLS semantic relations from narrative text in PubMed. Our method analyzes biomedical narrative text to collect semantic entity pairs, and extracts multiple semantic, syntactic and orthographic features for the collected pairs. We experiment with seeded k-means clustering with various distance metrics. We create and annotate a ground truth corpus according to the top two levels of the UMLS semantic relation hierarchy. We evaluate our system on this corpus and characterize the learning curves of different clustering configuration. Using KL divergence consistently performs the best on the held-out test data. With full seeding, we obtain macro-averaged F-measures above 70% for clustering the top level UMLS relations (2-way), and above 50% for clustering the second level relations (7-way).

Introduction
Biomedical documents are abundant in relations between concepts. For example, the sentence “The long-chain n-3 polyunsaturated fatty acids have cardioprotective effects, which may be partly due to their anti-inflammatory properties,” states at least two relations: the “long-chain n-3 polyunsaturated fatty acids” produces¹ “cardioprotective effects”; the “cardioprotective effects” are result_of “their anti-inflammatory properties”. However, such information, when locked in the narrative text, cannot be understood by computers due to lack of structure.

Mining relations from narrative text and making them accessible through a structured representation can benefit many studies, e.g., drug-drug, and drug-disease interaction studies. To this end, the Unified Medical Language System (UMLS) Semantic Network [1] can be used as a guide to align extracted structures. The network consists of the following:
- A set of semantic types, which provides a categorization of all concepts represented in the UMLS Metathesaurus®.
- A hierarchy of semantic relations, between semantic types.
This hierarchy has been developed and populated by hand in a top-down manner and undergoes periodic manual revisions [2]. As a result, collecting annotated relation instances and discovering those instances potentially characterizing new relations needs human annotation, which is labor intensive and time consuming. Our goal is to build a system that can expedite this process by mining relation instances from biomedical narrative text with little human intervention.

For this purpose, we need to be able to automatically identify the semantic relations between biomedical named entities. This is a nontrivial task as a semantic relation can be expressed in different ways using verbs (e.g., causes), prepositions (e.g., due to), or nouns (e.g., result of). For example, to say a symptom is the result_of a disease, one can use “due to”, “caused by”, or “result of”. In addition, solely relying on keywords themselves can be problematic because a given word can be polysemantic. For example, the word “undergo” in “patient undergoes homeopathy procedure” indicates that the patient is treated_by a treatment. The same word in “patient undergoes a severe seizure” means that the patient has_occurrence a disease/symptom. Context of the word “undergo” is necessary to characterize the relations between the named entities.

The UMLS semantic relations are organized in a hierarchy. For example, the relations disrupts, prevents, and complicates are categorized under a more generic relation affects. We focus on the top two levels of this hierarchy and test our method for identifying relations at both levels. In the rest of this paper, we first review related work, then describe the construction of our corpus. After that, we explain our automatic relation extraction method in detail, and present experimental studies.

Related Work
Semantic relation extraction from biomedical narrative text is an active area of research. Rosario et al. [3] compared five graphical models and neural networks in classifying seven relations between diseases and treatments, where the neural network outperformed all graphical models. Plake et al. [4] used finite automata to learn from training samples. Khoo et al. [5] identified causal relations with manually created syntactic patterns from MEDLINE [6]. Sibanda et al. [7] used support vector machines (SVM) [8] to recognize disease-treatment relations in discharge summaries.

¹ Italic font in the main paper denotes the matching relations in the UMLS Semantic Network [1].
Clinical NLP systems such as MedLEE [9] and SemRep [10] apply hand-crafted syntactic and semantic rules to extract UMLS semantic relations. Co-occurrence patterns in MEDLINE [6] have also been explored to identify gene and protein synonyms [11], protein-protein interactions [12] etc. (see [13] for a review). Recently, semi-supervised or unsupervised acquisition of semantic relations has gained traction in the general NLP domain, where the methods typically include clustering and co-clustering algorithms that are often augmented with seeding or subsequent supervised classification [14][15][16]. We believe that these new developments towards demanding less annotated gold standard can shed light on the biomedical domain, where extracting the UMLS semantic relations largely depends on supervised learning.

### Data Preparation

Our data set consists of the biomedical abstracts from the PubMed database [17]. We obtained the data set by crawling medical abstracts from the PubMed database that were returned in response to the query term “clinic”. This query term was used to include a broad range of topics across the abstracts. We collected semantic entities mentioned in an abstract by applying the UMLS TFA parser [18] and extracting noun phrases from its phrase chunking output. We treated each noun phrase as a semantic entity and paired all phrases in one sentence. We filtered candidate semantic entity pairs based on whether a relation can exist between the semantic types of involved entities according to the UMLS Semantic Network. We focused on only the relations that are explicitly stated in the text. Two annotators, who have information science background and have completed college-level biology courses, annotated relations for each record. The annotators were presented with candidate semantic entity pairs and selected the best matching relations by following the UMLS semantic relation definitions. We found that some semantic relations in the third and fourth levels of the UMLS Semantic Network were either absent or were poorly represented in our corpus. Therefore, we limited ourselves to the seven relations in the top two levels of the UMLS Semantic Network. We performed double annotation for each semantic entity pair. The annotation lasted three months, covered 207 medical abstracts (3002 sentences) and produced 10082 semantic entity pairs. The initial Kappa statistic for inter-annotator agreement is 0.81 reflecting high agreement [19]. The annotators then discussed on disagreements and were able to resolve most of them. We discarded 124 pairs with irresolvable disagreements. The number of instances of each semantic relation in the gold standard data set is listed in Table 1.

### Methods

We build a system that can automatically group UMLS concept pairs from biomedical narrative text into clusters where the grouping largely corresponds to the current semantic relation classes. We experiment with the k-means clustering framework under different configurations of distance metrics and seeding.

| Relation                                | Count |
|-----------------------------------------|-------|
| Associated_with (AW)                    | 9561  |
| Spatially_related_to (SRT)              | 488   |
| Functionally_related_to (FRT)           | 4719  |
| Conceptually_related_to (CRT)           | 3177  |
| Physically_related_to (PRT)             | 506   |
| Temporally_related_to (TRT)             | 497   |
| Isa (ISA)                               | 397   |

Table 1 Semantic Relation distribution. AW and ISA are top level UMLS relations, the rest are in the second level. Note that there are 174 AW instances that do not fall in the second level *RT relations.

Figure 1 shows the workflow of our system. We use the UMLS TFA parser [18] to perform tokenization, part-of-speech tagging, and phrase chunking on narrative sentences. The phrases identified by the UMLS TFA parser constitute “minimal syntactic units” and consist of lexical elements. A lexical element can be a single-word term, or a multi-word term if that term is determined to be an independent unit in general English or medical dictionaries/thesauri (e.g., MeSH [20] and the UMLS SPECIALIST lexicon [21]). The TFA parser then applies semantic-syntactic rules over lexical elements to chunk them into phrases. For the resultant noun phrases, we extract their semantic types using UMLS MetaMap Transfer (MMTx) [22].

![Figure 1 System workflow](image)
cept” for “the high rate”. This is motivated by the fact that certain semantic relation preferentially holds between specific sets of semantic types and vice versa.

- Lexical features that include all words in a sentence except for stop words. For example “high” and “rate” in the phrase “the high rate” in Figure 2. The intuition is that the words will help further distinguish semantic entities, in addition to semantic categories.

- Orthographic features that include punctuation, capitalization and the presence of digits. For example, capitalized phrase can be a proper name, often an instance of some disease, symptom etc. and likely to be in an is_a relation.

- Statistical features that include phrase length, sentence length and distance between phrases, all counted in terms of words. The intuition is that the relative distance between phrases can help differentiate semantic relations.

- Part-of-speech tags such as “verb” for “stain” and “noun” for “stain”, which help distinguish that the word indicates a relation or is part of an entity.

- Syntactic features that include the syntactic links between semantic entities. Intuitively, similar link paths may indicate similar semantic relations, such as the two example link paths in Figure 2.

We next give more details on syntactic features. We use the Link Parser [23] to extract syntactic links between the words in a sentence. The Link Parser identifies 106 types of syntactic links and associates words with left and right connectors. A pair of compatible connectors forms a link. To adapt the Link Parser output for our task, we convert word links to phrase level links by performing the following steps: if the words at both ends of a link are in two different phrases, then that link is regarded as an inter-phrase link and is retained; if the words on both ends of a link are in one phrase, then the link is treated as an intra-phrase link and is discarded. For example, Figure 2 shows the result of applying the above procedures on the sentence “Further studies with more samples are needed in order to explain the high rate found among the pediatric patients in this research study”. In Figure 2, we color multi-word phrases as blue and inter-phrase links as red. Other links are discarded.

We observe that the link labels often have semantic implications that are useful in characterizing the relations between the connected semantic entities. To explore this observation, we generalize Sibanda’s syntactic n-grams [7]. We divide phrase-level links into introductory links, intermediate links, and closing links, according to whether they are before, between, or after the entity pairs. For example, in Figure 2, the semantic entity pairs “the high rate”, “the pediatric patients” (P1) and “the high rate”, “this research study” (P2) share the relation occurs_in. These two pairs have common link types as well. Tracing the intermediate link path for the pair P2 as highlighted with green in Figure 2, the intermediate links include “Mvp” (indicating participle modifiers), “MVp” (connecting verbs to modifying preposition phrases), and “Js” (connecting prepositions to their objects). A similar analysis shows that the pair P1 shares the same intermediate links (highlighted with yellow in Figure 2) once ignoring subscripts2, suggesting that similar relation holds here as in P2.

We construct link bigrams for all three types of links. Motivated by the observation that longer link span may lead to weaker phrasal relation, we also add the link span (defined as 1 + #phrases between link endpoints) as a feature.

![Figure 2 Link grammar output of an example sentence.](image)

**Figure 2** Link grammar output of an example sentence. In this example, multi-word phrases are highlighted with blue color and inter-phrase links are highlighted with red color. Two example link paths are also highlighted with green and yellow respectively.

### Clustering Semantic Relations

We use the k-means clustering algorithm as we already know the number of clusters. Denote the data set as $Y = \{y_1, ..., y_k\}$, we want to form $k$ disjoint clusters $\hat{Y} = \{\hat{y}_1, ..., \hat{y}_k\}$, that is, $\bigcup_{i=1}^{k} \hat{y}_i = Y$, and $\forall i \neq j$, $\hat{y}_i \cap \hat{y}_j = \emptyset$. Let $X = \{x_1, ..., x_w\}$ be the features. These features are transformed into appropriate distance metrics between data points, which guide the formation of clusters by k-means. We use the Gmeans package [24], which minimizes an aggregated measure of intra-cluster distances called incoher-

---

2 Subscripts are used to encode fine grammatical constraints. For example, the “p” is a subscript in “MVp”, indicating prepositional modifying phrases to verbs.
ence. We experiment on the seeded k-means with several distance metrics including the Euclidean distance, the cosine similarity, and the Kullback-Leibler (KL) divergence. Let $c_j$ be the center of the cluster $j$, which is computed by averaging across all data points in that cluster. The Euclidean distance incoherence is

$$\mathcal{E}((\hat{y}_j)_{j=1}^k) = \sum_{j=1}^k \sum_{y \in \hat{y}_j} (y - c_j)^T(y - c_j).$$

The cosine similarity incoherence is calculated as

$$\mathcal{Q}((\hat{y}_j)_{j=1}^k) = \sum_{j=1}^k \sum_{y \in \hat{y}_j} y^Tc_j.$$

The KL divergence incoherence indicates the information loss due to clustering and is formulated as

$$\mathcal{D}((\hat{y}_j)_{j=1}^k) = \sum_{j=1}^k \sum_{y \in \hat{y}_j} p(y)KL(p(X|y), p(X|\hat{y}_j))$$

where $X, y, \hat{y}_j$ are viewed as random variables.

**Experimental Results and Discussions**

We evaluate our system on the top two levels of the UMLS Semantic Network. To test the generalizability of our semi-supervised clustering, we split our corpus into a training set and a testing set at a 8:2 ratio, stratified by semantic relation types at the second level (top level then automatically stratified). For each distance metric and seeding configuration, we run k-means clustering 30 times to obtain statistically robust results. For each run, we randomly draw seeds at specified fraction. We evaluate performance by creating a confusion matrix: we assign cluster labels so that we can obtain the confusion matrix with the strongest diagonal [25]. We then compute per-class as well as micro- and macro- averaged precision, recall and F-measure, which are common clustering evaluation metrics [26]. Let TP denote the number of true positives, FP denote the number of false positives and FN denote the number of false negatives, the definition of precision is $P = TP/(TP + FP)$, recall is $R = TP/(TP + FN)$, F-measure is $F = 2 \times P \times R / (P + R)$. For each distance metric-seeding configuration, we report averaged results over the 30 runs. We find that the KL divergence consistently gives the best F-measures for varying seed fractions.

Figure 3 shows the learning curves of the seeded clustering with KL divergence as the distance metric. For both levels of the UMLS semantic relation hierarchy, the performance keeps improving beyond 50% seeding, but with decreasing speed. Due to space limitation, we leave the results of the other two distance metrics in the Appendix B, which were inferior to those of the KL divergence. This suggests that the seeded clustering is sensitive to the choice of the distance metric and that the KL divergence is a suitable distance metric on our dataset. See Appendix B for more comparisons.

![Figure 3 Macro-averaged and micro-averaged Precision, Recall and F-measure on 2-way and 7-way relation using KL divergence as the distance metric. Results are averaged over 30 runs, confidence intervals at $\alpha = 0.05$ are also shown, most of which are small, suggesting statistical stability.](image-url)
With full seeding on the training data, our method generates a 2-way macro-averaged F-measure above 70% and a 7-way macro-averaged F-measure above 50%. With half seeding, our method still achieves a 2-way macro-averaged F-measure above 65% and a 7-way macro-averaged F-measure around 45%. This result suggests that the demand for seeding can be relaxed and our method can be used to automatically identify relation instances and provide a reasonable starting point with pre-labels to facilitate the human annotation process.

We present detailed per-class evaluation on k-means algorithm using KL divergence with varying seed fractions (up to 50%\(^3\)) in Table 2 and Table 3. Not surprisingly, we see that the class imbalance took its toll on the clustering performance. For example, in the top level (Table 2), we have 9561 examples of the AW relation vs. 397 examples of the ISA relation. Precision and recall of ISA is much lower than AW. For the second level (Table 3), big performance drop is also seen in less populated classes such as AW\(^4\), PRT and TRT.

Table 2 Performance per class of k-means clustering using KL divergence with random seeds on the relations from the top level of the UMLS Semantic Network.

| Seed | Relation | Precision | Recall | F-measure |
|------|----------|-----------|--------|-----------|
| 10%  | AW       | 96.45%    | 98.83% | 97.62%    |
|      | ISA      | 30.78%    | 10.94% | 15.78%    |
| 20%  | AW       | 96.79%    | 98.83% | 97.80%    |
|      | ISA      | 42.62%    | 19.83% | 26.64%    |
| 30%  | AW       | 96.88%    | 98.61% | 97.74%    |
|      | ISA      | 41.09%    | 22.31% | 28.54%    |
| 40%  | AW       | 97.05%    | 98.50% | 97.77%    |
|      | ISA      | 42.84%    | 26.75% | 32.64%    |
| 50%  | AW       | 97.16%    | 98.55% | 97.85%    |
|      | ISA      | 45.81%    | 29.53% | 35.78%    |

Conclusion and Future Work

We presented a semi-supervised approach to automatically identify semantic relations according to the definitions in the UMLS Semantic Network. We created a corpus of semantic entity pairs whose relations were doubly annotated according to the top two levels of the UMLS semantic relation hierarchy. We demonstrated that our semi-supervised method has reasonable accuracy and coverage at both levels of resolution. By studying the learning curves of the seeded k-means with the KL divergence as the distance metric, we showed that the demand for seeding in the training data can be relaxed by half without greatly decreasing the performance. Therefore, our system can be used to assist with expert reviews on the semantic relation annotation task.

For future work, we note that our seeding is random and does not take into consideration how informative an example is. An active learning approach for picking the seeds could potentially further reduce the amount of required seeds and maintain similar levels of precision, recall, and F-measure.

Table 3 Performance per class of k-means clustering using KL divergence with random seeds (fractions 10% to 50%) on the relations from the second level of the UMLS Semantic Network. AW here includes AW instances that do not fall into *RT categories.

| Seed | Relation | Precision | Recall | F-measure |
|------|----------|-----------|--------|-----------|
| 10%  | AW       | 27.50%    | 8.43%  | 12.20%    |
|      | CRT      | 49.35%    | 55.02% | 51.93%    |
|      | FRT      | 57.92%    | 64.77% | 61.12%    |
|      | ISA      | 36.39%    | 15.68% | 21.47%    |
|      | PRT      | 21.04%    | 10.83% | 14.09%    |
|      | SRT      | 35.90%    | 21.31% | 26.39%    |
|      | TRT      | 22.38%    | 13.71% | 16.73%    |
| 20%  | AW       | 33.56%    | 16.96% | 21.79%    |
|      | CRT      | 54.12%    | 58.30% | 56.09%    |
|      | FRT      | 60.41%    | 67.25% | 63.63%    |
|      | ISA      | 43.84%    | 18.76% | 26.13%    |
|      | PRT      | 26.91%    | 15.17% | 19.21%    |
|      | SRT      | 44.19%    | 30.79% | 36.05%    |
|      | TRT      | 26.21%    | 18.61% | 21.50%    |
| 30%  | AW       | 34.11%    | 16.76% | 21.54%    |
|      | CRT      | 56.80%    | 59.90% | 58.28%    |
|      | FRT      | 62.63%    | 68.56% | 65.44%    |
|      | ISA      | 47.86%    | 25.09% | 32.65%    |
|      | PRT      | 30.22%    | 19.57% | 23.57%    |
|      | SRT      | 50.84%    | 38.69% | 43.78%    |
|      | TRT      | 28.47%    | 23.20% | 25.35%    |
| 40%  | AW       | 36.77%    | 20.39% | 25.41%    |
|      | CRT      | 59.03%    | 63.00% | 60.94%    |
|      | FRT      | 64.05%    | 68.99% | 66.42%    |
|      | ISA      | 52.89%    | 29.44% | 37.65%    |
|      | PRT      | 34.19%    | 22.57% | 27.01%    |
|      | SRT      | 51.68%    | 43.30% | 47.01%    |
|      | TRT      | 29.86%    | 23.98% | 26.47%    |
| 50%  | AW       | 36.15%    | 23.43% | 28.01%    |
|      | CRT      | 60.55%    | 65.17% | 62.77%    |
|      | FRT      | 66.11%    | 69.55% | 67.78%    |
|      | ISA      | 55.49%    | 33.80% | 41.86%    |
|      | PRT      | 37.13%    | 24.73% | 29.58%    |
|      | SRT      | 55.04%    | 49.97% | 52.26%    |
|      | TRT      | 32.47%    | 27.28% | 29.54%    |

\(^3\) Results for seed fraction above 50% are shown in Appendix A.

\(^4\) In the second level, AW refers to those 174 instances that do not fall into *RT relations.
References

1. UMLS Semantic Network
   http://www.nlm.nih.gov/pubs/factsheets/umlssemn.html
2. McCray AT. An upper level ontology for the biomedical domain. Comp Funct Genom 2003; 4:80-4.
3. Rosario B, Hearst M. Classifying Semantic Relationships in BioScience Text. ACL 2004.
4. Plake C, Schiemann T, Pankalla M, Hakenberg J, Larser U. Ali Baba: PubMed as a graph. Bioinformatics, 22(19): 2444-2445, 2006
5. Khoo C, Chan S, Niu Y. Extracting Causal Knowledge from a Medical Database Using Graphical Patterns. ACL 2000. pp. 336-343.
6. MEDLINE
   http://www.nlm.nih.gov/pubs/factsheets/medline.html
7. Sibanda T. Was the Patient Cured? Understanding Semantic Categories and Their Relationships in Patient Records. Master Thesis, MIT, 2006.
8. Vapnik V. The Nature of Statistical Learning Theory. Berlin : Springer-Verlag, 1995.
9. Hristovski, D., Friedman, C., Rindflesch, T. C., & Peterlin, B. Exploiting semantic relations for literature-based discovery. AMIA 2006
10. Rindflesch, T and Marcelo, F. The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. Journal of biomedical informatics 36.6 (2003): 462-477.
11. Cohen, A. M., Hersh, W. R., Dubay, C., & Spackman, K. (2005). Using co-occurrence network structure to extract synonymous gene and protein names from MEDLINE abstracts. BMC bioinformatics, 6(1), 103.
12. Bunescu, R., Mooney, R., Ramani, A., & Marcotte, E. (2006, June). Integrating co-occurrence statistics with information extraction for robust retrieval of protein interactions from Medline. In Proceedings of the Workshop on Linking Natural Language Processing and Biology: Towards Deeper Biological Literature Analysis (pp. 49-56). Association for Computational Linguistics.
13. Cohen, T., & Widdows, D. (2009). Empirical distributional semantics: methods and biomedical applications. Journal of Biomedical Informatics, 42(2), 390-405.
14. Bollegala, T, Yutaka M, and Mitsuru I. "Relational duality: Unsupervised extraction of semantic relations between entities on the web." Proceedings of the 19th international conference on World wide web. ACM, 2010.
15. Sun, A, Ralph G, and Satoshi S. "Semi-supervised Relation Extraction with Large-scale Word Clustering." ACL. 2011.
16. Mohamed, T, Estevam H, and Tom M. "Discovering relations between noun categories." Proceedings of EMNLP, 2011.
17. PubMed. NCBI.
   http://www.ncbi.nlm.nih.gov/sites/entrez?db=pubmed
18. Text Tools from Lexical System Group.
   http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/textTools/current/Usages/Parser.html
19. Fleiss, J.L., 1981. Statistical Methods for Rates and Proportions, Wiley.
20. MeSH: http://www.ncbi.nlm.nih.gov/mesh
21. Specialist Lexicon.
   http://lexsrv3.nlm.nih.gov/LexSysGroup/index.html
22. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. AMIA 2001.
23. Sleator D, Temperley D. Parsing english with a link grammar. Technical Report, Carnegie Mellon University. 1991.
24. Dhillon I, Guan Y. Information-Theoretic Clustering of Sparse Co-Occurrence Data. UTCS Technical Report #TR-03-39.
25. Dhillon IS, Guan Y. Information Theoretic Clustering of Sparse Co-Occurrence Data. Proceedings of the Third IEEE International Conference on Data Mining, p.517, November 19-22, 2003.
26. Manning, C.D. et al. 2008. Introduction to information retrieval. Cambridge University Press Cambridge.
Appendix A

This section continues the per-class evaluation for k-means using KL divergence. Table 4 shows the continuation of Table 3 with seed fraction increasing to 100%. Table 5 shows the continuation of Table 2 with seed fraction increasing to 100%.

**Table 4** Performance per class of k-means clustering using KL divergence with random seeds (fractions 60% to 100%) on the relations from the second level of the UMLS Semantic Network. See Table 1 for notation on abbreviations.

| Seed | Relation | Precision | Recall | F-measure |
|------|----------|-----------|--------|-----------|
| 60%  | AW       | 39.19%    | 25.69% | 30.38%    |
|      | CRT      | 62.59%    | 66.51% | 64.48%    |
|      | FRT      | 67.42%    | 70.53% | 68.93%    |
|      | ISA      | 55.60%    | 35.47% | 43.18%    |
|      | PRT      | 40.10%    | 28.87% | 33.46%    |
|      | SRT      | 58.65%    | 56.80% | 57.57%    |
|      | TRT      | 34.28%    | 28.74% | 31.16%    |
| 70%  | AW       | 40.30%    | 27.06% | 31.80%    |
|      | CRT      | 63.34%    | 68.07% | 65.61%    |
|      | FRT      | 68.46%    | 70.49% | 69.45%    |
|      | ISA      | 56.28%    | 36.71% | 44.36%    |
|      | PRT      | 39.91%    | 29.17% | 33.64%    |
|      | SRT      | 58.41%    | 58.45% | 58.35%    |
|      | TRT      | 33.87%    | 29.46% | 31.44%    |
| 80%  | AW       | 35.76%    | 26.08% | 29.75%    |
|      | CRT      | 64.35%    | 69.26% | 66.71%    |
|      | FRT      | 70.13%    | 70.96% | 70.54%    |
|      | ISA      | 56.96%    | 39.19% | 46.36%    |
|      | PRT      | 42.92%    | 32.83% | 37.16%    |
|      | SRT      | 61.37%    | 65.22% | 63.17%    |
|      | TRT      | 34.92%    | 31.09% | 32.86%    |
| 90%  | AW       | 38.06%    | 28.33% | 31.95%    |
|      | CRT      | 65.31%    | 70.87% | 67.97%    |
|      | FRT      | 71.43%    | 71.20% | 71.31%    |
|      | ISA      | 58.41%    | 41.15% | 48.26%    |
|      | PRT      | 44.52%    | 35.70% | 39.60%    |
|      | SRT      | 62.01%    | 67.66% | 64.68%    |
|      | TRT      | 34.89%    | 31.33% | 33.00%    |
| 100% | AW       | 34.48%    | 29.41% | 31.75%    |
|      | CRT      | 66.09%    | 72.24% | 69.03%    |
|      | FRT      | 72.71%    | 71.40% | 72.05%    |
|      | ISA      | 58.93%    | 42.31% | 49.25%    |
|      | PRT      | 45.00%    | 36.00% | 40.00%    |
|      | SRT      | 63.89%    | 71.13% | 67.32%    |
|      | TRT      | 35.87%    | 33.67% | 34.74%    |

**Table 5** Performance per class of k-means clustering using KL divergence with random seeds on the relations from the top level of the UMLS Semantic Network. See Table 1 for notation on abbreviations.

| Seed | Relation | Precision | Recall | F-measure |
|------|----------|-----------|--------|-----------|
| 60%  | AW       | 97.29%    | 98.62% | 97.95%    |
|      | ISA      | 49.54%    | 32.78% | 39.28%    |
| 70%  | AW       | 97.39%    | 98.46% | 97.92%    |
|      | ISA      | 48.66%    | 35.47% | 40.94%    |
| 80%  | AW       | 97.43%    | 98.48% | 97.95%    |
|      | ISA      | 49.59%    | 36.41% | 41.95%    |
| 90%  | AW       | 97.49%    | 98.36% | 97.92%    |
|      | ISA      | 48.77%    | 38.16% | 42.79%    |
| 100% | AW       | 97.55%    | 98.32% | 97.94%    |
|      | ISA      | 49.21%    | 39.74% | 43.97%    |
Appendix B

Figure 4 shows the learning curves of using the cosine similarity as the distance metric in clustering semantic entity pairs according to the top two levels of UMLS semantic relations. Figure 5 shows the corresponding learning curves for using the Euclidean distance as the distance metric. Comparing them to Figure 3, it can be seen that the cosine similarity and

the Euclidean distance are consistently outperformed by the KL divergence distance metric when cluster seeds are given. Moreover, evaluation metrics on the cosine similarity and the Euclidean distance often have larger statistical variations (bigger confidence intervals) than those on the KL divergence. This further suggests that KL divergence as the distance metric tends to be more statistically stable on our dataset.

![Figure 4](image1.png)

![Figure 5](image2.png)

Figure 4 Macro-averaged and micro-averaged Precision, Recall and F-measure on 2-way and 7-way relation clustering using cosine similarity as distance. Results are averaged over 30 runs, confidence intervals at α=0.05 are also shown.
Figure 5 Macro-averaged and micro-averaged Precision, Recall and F-measure on 2-way and 7-way relation clustering using the Euclidean distance. Results are averaged over 30 runs, confidence intervals at $\alpha=0.05$ are also shown.