Investigation of Unsupervised Pattern Learning Techniques for Bootstrap Construction of a Medical Treatment Lexicon

Rong Xu, Alex Morgan, Amar K Das
Biomedical Informatics Program
Stanford University
Stanford, CA 94305, USA
xurong@stanford.edu

Alan Garber
Primary Care and Outcomes Research
Stanford University
Stanford, CA94305, USA

Abstract

Dictionaries of biomedical concepts (e.g. diseases, medical treatments) are critical source of background knowledge for systems doing biomedical information retrieval, extraction, and automated discovery. However, the rapid pace of biomedical research and the lack of constraints on usage ensure that such dictionaries are incomplete. Focusing on medical treatment concepts (e.g. drugs, medical procedures and medical devices), we have developed an unsupervised, iterative pattern learning approach for constructing a comprehensive dictionary of medical treatment terms from randomized clinical trial (RCT) abstracts. We have investigated different methods of seeding, either with a seed pattern or seed instances (terms), and have compared different ranking methods for ranking extracted context patterns and instances. When used to identify treatment concepts from 100 randomly chosen, manually annotated RCT abstracts, our medical treatment dictionary shows better performance (precision:0.40, recall: 0.92 and F-measure: 0.54) over the most widely used manually created medical treatment terminology (precision: 0.41, recall: 0.52 and F-measure: 0.42).

1 Introduction

Dictionary based natural language processing systems have been widely used in recognizing medical concepts from free text. For example, the MetaMap program is used to map medical text to concepts from the most widely used biomedical terminology, the Unified Medical Language System (UMLS) Metathesaurus (Aronson, 2000). It identifies various forms of UMLS concepts in text and returns them as a ranked list using a five-step process: identifying simple noun phrases (NP's), generating variants of each phrase, finding matched phrases, assigning scores to matched phrases and composing mappings. However, its performance largely depends on the quality of the underlying UMLS Metathesaurus and its manually created rules and variants. One study has shown that, of the medical concepts identified by human subjects, more than 40% were not in UMLS (Pratt, 2003). Other examples of mapping text to controlled biomedical terminologies include (Cohen, 2005) and (Fang, 2006). Many other systems make heavy use of biomedical terminologies directly such as the work of Blaschke, et al. (Blaschke, 2002) and Friedman et al. (Friedman, 2001).

Biomedical terminology is highly dynamic, both because biomedical research is itself highly dynamic, but also because there are essentially no constraints on the use of new terminological variants, making the terms used in free text quite different from the canonical forms listed in controlled terminologies. To contrast UMLS with actual text mentions, there are 150 different *chemotherapy* concepts in UMLS. The majority of these terms derive from the diseases they are used to treat. For example *cancer chemotherapy*, *AIDS chemotherapy*, *brain disorder chemotherapy*, and *alcoholism chemotherapy*. On the other hand, we have identified more than 1,000 different *chemotherapy* types mentioned in RCT (Randomized Clinical Trial) report abstracts, with most of the names derived

Proceedings of the Workshop on BioNLP, pages 63–70,
Boulder, Colorado, June 2009.

from the chemicals contained in the chemotherapy regimen, such as *platinum-based chemotherapy* or *fluorouracil-based chemotherapy*. There is little overlap between the *chemotherapy* terms in UMLS and the ones used in RCT abstracts. Even for simple drug names as *5-fluorouracil* and *tamoxifen*, there are many clinically distinct and important variants of these drugs which are absent in UMLS as distinct terms/concepts, such as *intralesional 5-fluorouracil*, *topical 5-fluorouracil*, *intrahepatic arterial 5-Fluorouracil*, *adjuvant sequential tamoxifen*, and *neoadjuvant tamoxifen*.

There has been considerable work on expanding the coverage of biomedical dictionaries through morphological variants, but these approaches require an initial term dictionary with reasonable extensive coverage. Examples include the approaches developed by Krauthammer and Nenadic (Krauthammer, 2004), Tsuruoka and Tsujii (Tsuruoka, 2004) & (Tsuruoka, 2003), Bodenreider, et al. (Bodenreider, 2002), and Mukherjea and colleagues (Mukherjea, 2004). An important shortcoming with static, human derived terminologies that cannot easily be addressed by looking for variants of existing terms is the fact that continual developments in medical therapies constantly gives rise to new terms. Examples include, *Apomab*, *Bapineuzumab*, *Bavituximab*, *Etaracizumab*, and *Figitumumab*. These all represent a new generation of targeted biological agents currently in clinical trials none of which appear in UMLS. Clearly we need to develop techniques to deal with this dynamic terminology landscape.

MEDLINE is the most extensive and authoritative source of biomedical information. Large quantities of biomedical text are available in MEDLINE's collection of RCT reports with over 500,000 abstracts available. RCT reports are a critical resource for information about diseases, their treatments, and treatment efficacy. These reports have the advantage of being highly redundant (a disease or treatment name is often reported in multiple RCT abstracts), medically related, coherent in writing style, trustworthy and freely available.

In our recent study (Xu, 2008), we have developed and evaluated an automated, unsupervised, iterative pattern learning approach for constructing a comprehensive disease dictionary from RCT abstracts. When used to identify disease concepts from 100 manually annotated clinical abstracts, the disease dictionary shows significant performance improvement (F1 increased by 35-88%) over UMLS and other disease terminologies. It remained to be demonstrated that these bootstrapping techniques are indeed rapidly retargetable and can be extended to other situations, and so we have extended our scope to investigate medical treatment names in addition to disease terms in this work.

Our approach is inspired by the framework adopted in several bootstrapping systems for learning term dictionaries, including (Brin, 1998), (**?**), and (Agichtein, 2000). These approaches are based on a set of surface patterns (Hearst , 1992), which are matched to the text collection and used to find instance-concept relations. Similar systems include that of Snow and colleagues (Snow, 2005), which integrates syntactic dependency structure into pattern representation and has been applied to the task of learning instance-of relations, and the approach developed of Caprosaso, et al. (Caprosaso, 2007) which focussed on learning text context patterns to identify mentions of point mutations.

All iterative learning systems suffer from the inevitable problem of spurious patterns and instances introduced in the iterative process. To analyze different approaches to addressing this issue, we have compared three different approaches to ranking extracted patterns and three different approaches to ranking extracted instances. Because such systems also depend on an initial seeding with either a seed pattern or term instance, an important question is whether these different starting points lead to different results. We investigated this issue by starting from each point separately and compared the final results.

2 Data and Methods

2.1 Data

509,308 RCT abstracts published in MEDLINE from 1965 to 2008 were parsed into 8,252,797 sentences. Each sentence was lexically parsed to generate a parse tree using the Stanford Parser. The Stanford Parser (Klein, 2003) is an unlexicalized natural language parser, trained on a non-medical document collection (Wall Street Journal). We used

the publicly available information retrieval library, Lucene, to create an index on sentences and their corresponding parse trees. For evaluation and comparison, 241,793 treatment terms with treatment related semantics types from UMLS were used.

2.2 Unsupervised Instance Extraction and Pattern Discovery

Figure 1 describes the bootstrapping algorithm used in learning instances of treatment and their associated text patterns. The algorithm can operate in two modes, either starting with a seed pattern p_0, which represents a typical way of writing about treatments, or a set of seed instances, (d_i). For example, the seed pattern we used was "*treated with NP*" (*NP*: noun phrase). The program loops over a procedure consisting of two steps: instance extraction and pattern discovery. In the instance extraction step, patterns are used as search queries to the local search engine. The parse trees with given patterns are retrieved and noun phrases (instances of treatments) following the pattern are matched from the parse trees. In the pattern discovery step, instances extracted from the previous iteration are used as search queries to the local search engine. Corresponding sentences containing instance mentions are retrieved and the bigrams (two words) in front of instances are extracted as patterns. When seeding with an initial pattern, only two iterations are typically needed, as experience shows that most of reliable patterns and instances have been discovered at this stage. The algorithm stops after a single iteration when seeding with a list of instances.

2.3 Selecting Seed Instances

Of the 241,793 treatment related terms in the UMLS, only about 22,000 (9%) of these have appeared in MEDLINE RCT reports. We randomly selected 500 drug terms and 500 medical procedure terms from the 22,000 terms as seed instances and used them in the pattern discovery system described above.

2.4 Pattern Ranking

A newly discovered pattern is scored on how similar its output (instances associated with the pattern) is to the output of the initial seed pattern. Intuitively, a reliable pattern is one that is both highly

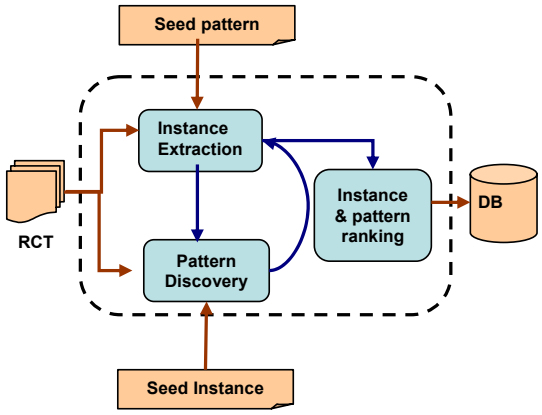

Figure 1: General scheme of the iterative method.

precise (high precision) and general (high recall). Using the output instances from the seed pattern p_0 as a comparison, we developed *Precision Based*, *Recall Based*, and *F1 Based* algorithms to rank patterns. We define $instances(p)$ to be the set of instances matched by pattern p, and the intersection $instances(p) \bigcap instances(p_0)$ as the set of instances matched by both pattern p and p_0.

1. *Precision Based* rank:

$$score1(p) = \frac{instances(p) \bigcap instances(p_0)}{instances(p)} \quad (1)$$

   The *precision based* ranking method favors specific patterns.

2. *Recall Based* rank:

$$score2(p) = \frac{instances(p) \bigcap instances(p_0)}{instances(p_0)} \quad (2)$$

   The *recall based* ranking method favors general patterns.

3. *F1 based* rank:

$$score3(p) = \frac{2 \times score1(p) \times score2(p)}{score1(p) + score2(p)} \quad (3)$$

   A combination of the *Precision Based* and the *Recall Based* evaluation methods is the *F1*

Based ranking method, which takes into account both pattern specificity and pattern generality. This method favors general patterns while penalizing overly specific patterns.

2.5 Instance Ranking

A reliable instance is one that is associated with a reliable pattern many times. We experimented with three ranking algorithms:

1. *Abundance Based* rank: A treatment instance(d) that is obtained multiple times is more likely to be a real treatment concept when compared with one that has only a single mention in the whole corpus. We define $scoreA(d)$ as the number of times where d appears in the corpus.

2. *Pattern Based* rank: A treatment instance obtained from multiple patterns is more likely to be a real treatment concept when compared with the one that was obtained by a single pattern (p). *Pattern Based* rank takes into account the number of patterns that generated the instance, score of those patterns, and the number of times that the instance is associated with each pattern ($count(p, d)$).

$$scoreB(d) = \sum_{i=0}^{n} \log score3(p_i) \times count(p_i, d) \quad (4)$$

3. *Best Pattern Based* rank: A treatment instance obtained from a highly ranked pattern is more likely to be a real treatment concept when compared with the one that was obtained from a poorly ranked pattern. First the instances are ranked by the best pattern (p_b) that generated the instances and then further ties are broken by the number of times the instance is associated with that pattern ($count(p, d)$) to provide $scoreC(d)$.

2.6 Comparison of Patterns Derived from Different Seed Types

The patterns extracted when starting with either seed instances or a seed pattern are ranked by the *recall based* method and *F1-based* method, then the overlaps at different cutoffs are measured to assess the similarity of the patterns discovered by starting with the different starting seed types.

2.7 Evaluation of Stanford Parser in Identifying Treatment Noun Phrase

An important question is how accurate the Stanford Parser is at identifying the relevant term boundaries. We used manually curated treatment names from UMLS to measure the accuracy of the Stanford Parser in identifying treatment noun phrases. With $NPcount(treatment)$ defined as number of times that the Stanford Parser identifies a treatment as noun phrase or part of a noun phrase in the data and $count(treatment)$ as number of times the treatment appears in the data.

$$accuracy = \frac{1}{n} \sum_{i=0}^{n} \left(\frac{NPcount(d_i)}{count(d_i)} \right) \quad (5)$$

2.8 Evaluation of the extracted treatment lexicon

We assessed the quality (precision and recall) of our lexicon by using it to identify treatment concepts in 100 randomly selected RCT abstracts where treatment names were manually identified. In addition, we also compared the performance of our lexicon with that of UMLS.

3 Results

3.1 Evaluation of Stanford Parser in Identifying Treatment Noun Phrases

Even though the Stanford Parser is trained on non-medical data, it is highly accurate in identifying treatments as noun phrases or parts of a noun phrase with accuracy of 0.95. The reason may be that medical treatments are indeed often noun phrases or parts of a noun phrase in RCT reports, and there are strong syntactical signals for their phrasal roles in the sentences. For example, treatments are often either the object of a preposition (e.g. *efficacy of fluorouracil* and *treated with fluorouracil*) or the subject of a sentence (e.g. *fluorouracil is effective in treating colon cancer*).

3.2 Comparison between Seed Types

There is considerable overlap in discovered patterns between starting with a single seed pattern and start-

ing with the 1,000 seed instances and little difference in overall performance. 12,241 patterns are found to be associated with the 1,000 seed treatment instances. However, only the most highly ranked patterns are relevant (see Evaluation of The Extracted Treatment Lexicon, below). Table 1 shows the intersection of the top ranked patterns between both seeding methods at different rank cut-offs. We find a very high level of intersection between the top ranked patterns from both initial seed types, for example eighteen of the top twenty patterns are identical. These results indicate that starting from either seed type leads to very similar results.

Rank	Recall Based	F1 Based
10	0.90	0.80
20	0.90	0.90
30	0.87	0.80
40	0.83	0.85
50	0.84	0.82
60	0.82	0.85
70	0.82	0.79
80	0.83	0.84
90	0.84	0.83
100	0.82	0.83

Table 1: : The ratio of overlap in the top ranking patterns discovered by different seed types

3.3 Pattern Ranking

Similar to the results observed in our previous study (Xu, 2008), the *Precision Based* metric assigns high scores to very specific but not generalizable patterns such as "*lornoxicam versus*" (Table 2), which appears only once in the data collection, while the top 10 patterns based on the *Recall Based* and *F1 Based* rankings are typical treatment related patterns. When a different seed pattern "*efficacy of*" was used, the top 10 patterns were the same with a different rank ordering.

3.4 Instance ranking

Table 3 shows the top 10 suggested treatment names when using "*treated with*" as the initial seed pattern. The rank of a proposed treatment instance is determined by the different ranking methods: *Abundance Based*, *Pattern Based*, or *Best Pattern Based* ranking

#	Precision based	Recall based	F1 based
1	beta-blockers nor	treated with	treated with
2	lornoxicam versus	treatment with	treatment with
3	piroxantrone and	effects of	efficacy of
4	heparin called	efficacy of	effects of
5	anesthetics containing	dose of	dose of
6	antioestrogens and	doses of	doses of
7	markedly adsorb	suggest that	suggest that
8	recover following	study of	safety of
9	Phisoderm and	response to	response to
10	MitoExtra and	effect of	effect of

Table 2: Top 10 patterns with "*treated with*" as seed pattern

algorithms. None of the top 10 extracted phrases on the basis of *Abundance Based* or *Pattern Based* are actual treatment names. These two ranking methods assign high ranks to common, non-specific phrases. The *Best Pattern Based* ranking method correctly identifies specific treatment mentions, mainly because it reduces the likelihood of selecting irrelevant patterns.

#	Abundance based	Pattern based	Best pattern based
1	patients	patients	placebo
2	treatments	the treatment	chemotherapy
3	the treatments	treatments	radiotherapy
4	children	the use	tamoxifen
5	the effect	children	antibiotics
6	no significant differences	surgery	insulin
7	placebo	the patients	interferon
8	surgery	changes	surgery
9	the effects	women	corticosteroids
10	the study	use	cisplatin

Table 3: Top 10 treatments when using "*treated with*" as the seed pattern

3.5 Evaluation of the Extracted Treatment Lexicon

Our dictionary derived from using "*treated with*" as the seed pattern with two bootstrapping itera-

Count	Cutoff	Precision	Recall	F1
17,683	1.0%	0.404	0.921	0.540
88,415	5%	0.127	1.0	0.22
132,623	7.5%	0.105	1.0	0.187
176,832	10%	0.088	1.0	0.160

Table 4: Precision, recall and F1 at 4 cutoff values

tions consists of 1,768,320 candidate instances and 78,037 patterns, each with an accompanying confidence score. The top 20 patterns are associated with more than 90% of the instances. We evaluated the quality of the dictionary by using it to identify treatment concepts in 100 randomly selected abstracts where treatment names were manually annotated. There were an average of three treatment names per test abstract. Table 4 shows the precision, recall and F1 values when instances are ranked by the *best pattern based* ranking method (ScoreC). The precision, recall and F1 values at each cut-off (percentage of all instances) were averaged across the 100 abstracts.

The precision, recall and F1 of the UMLS Metathesaurus in identifying treatment names from the test dataset are 0.41, 0.52 and 0.42 respectively. The performance using UMLS on this task is consistent with a previous study (Pratt, 2003). The low precision may due to the fact that UMLS often tags irrelevant names as treatment related names. For example, common, non-specific terms such as *drug*, *agent*, *treatment* and *procedure* appear in the dictionary derived from UMLS. However, we chose not to edit the lexicon derived from UMLS as it is unclear how to do so in a systematic matter without essentially creating a new version of UMLS, and we are interested in studying methods that do not rely on any human involvement (our Discussion describes the possible inclusion of human judgments). Also, the low recall of UMLS is not surprising given the fact that the names specified in UMLS are often not the terms authors use in writing. The performance of our dictionary (precision: 0.40, recall: 0.92, F1: 0.54) is a dramatic improvement over using UMLS. Our recall is high since all the terms are learned from the literature directly and exemplify the manner in which authors write RCT reports. However, the precision of our dictionary is still low (see Discussion).

4 Discussion

We have demonstrated an automated, unsupervised, iterative pattern learning approach for bootstrapping construction of a comprehensive treatment lexicon. We also compared different pattern and instance ranking methods and different initial seed types (instances or patterns). On the task of term identification, use of our boostrapped lexicon increased performance over using the most widely used manually curated terminology (UMLS). We have extended our previous work to the identification of new terminology types, demonstrating the versatility of this approach. Our approach may also be used with other data sources such as general health related web pages. However, there is still significant space in which to seek improvement in increasing the coverage of our lexicon and the quality of our patterns.

Although useful in demonstrating the proof of concept and allowing us to examine different ranking methods, focusing on bigrams that precede noun-phrases limited the space of patterns that we could potentially examine. More complex patterns might be involved. For example, in the sentence "*Pravastatin is an effective and safe drug*" (PMID 08339527), there is a distinctive treatment related pattern "*NP is an effective and safe drug*" that our technique does not capture. However, most key terms are mentioned in multiple contexts. For example, *Pravastatin* appears with the seed pattern *treatment with* more than 200 times. As our corpus of literature increases, redundancy will increase the likelihood of a treatment term being matched by the type of patterns we recognize. The rapid growth of biomedical knowledge and literature, which makes our automatically generated medical treatment vocabulary necessary, can also act to increase its coverage over time.

In order to keep our algorithm simple, we did not perform deep grammatical analysis. For example, in the sentence "*Treatment of the subjects with atorvastatin decreased the abundance of IL-12p35 mRNA in mononuclear cells*" (PMID 12492458), *atorvastatin* is associated with *treatment of*, not *subjects with*. Since our algorithms simply extracts the two words in front of treatment names, *subjects with* will be extracted as treatment related pattern. In fact, *subjects with* is a disease related pattern in RCT reports, for

example "*34 subjects with asthma*". But our pattern ranking algorithm will assign a low score to *subjects with* since the terms associated with this pattern are more disease related and have little overlap with the output of the seed pattern *treatment with*.

Our instance ranking assigns high confidence scores to common and non specific terms like *this drug*, *the treatment* or *this procedure* since they are often associated with highly ranked patterns many times. These anaphoric terms often refer to treatment names previously specified. There are at least two ways to address this problem. The first is to assign low scores to terms starting with a determiner such as *the* or *this*. Another way to improve the instance ranking algorithm is to take into account of the overall context of the term. For example, these anaphora often appear in specific sections of RCT reports such as the *result* section, and refer to terms from previous sections. Specific examples include "*Treatment with this drug should be attempted in intractable cases*" (PMID 09038009) and "*The efficacy of the treatment was 88 and 95% in group 1 and 2, respectively*" (PMID 14520944). The terms from *title*, *background* or *conclusion* sections could be assigned higher scores than the ones from *result* section. Beyond these simple heuristics, more sophisticated approaches might take advantage of the work in anaphora resolution, such as (Baldwin, 2001).

The lexicon consists of terms with mixed hierarchies, including general terms as *chemotherapy*, *surgery*, *corticosteroids*, *antibiotics*, and specific terms as *fluorouracil*, *oral or intravenous 5-Fluorouracil*, *cisplatin*, *nephrectomy*. In order to make this dictionary more useful, additional work is needed to organize the terms and build ontologies based on the lexicon.

Previous work has shown that learning multiple semantic types simultaneously can improve precision (Thelen, 2002) & (Curran, 2007), and it remains to be seen if that approach can be combined with the prioritization of pattern and extracted instance rankings here to give better overall performance. Other possible extensions and improvements include various approaches to slow the learning process and discover new patterns and instances more conservatively, at the expense of more iterations. Further improvements can be expected from integrating active learning approaches to include the involvement of a human judge in the process, analogous to the tag-a-little, learn-a-little method proposed as part of the Alembic Workbench (Day, 1997). Because our approach ranks both extracted patterns and instances, it is amenable to such techniques. Indeed, active learning has been found to provide considerable gains in corpus annotation (Tomanek, 2007) & (Buyko, 2007), and can be a model for semi-automated terminology compilation.

All the data and code are available on request from the author.

Acknowledgments

RX is supported by NLM training grant LM007033 and Stanford Medical School.

References

E. Agichtein, L Gravano. 2000. *Snowball: extracting relations from large plaintext collections*, *In Proc of the 5th ACM conference on Digital libraries* .

A.R. Aronson 2001. *Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp*:17-21.

B. Baldwin 2001. *Text and knowledge mining for coreference resolution. Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*:1-8.

C. Blaschke, A. Valencia. 2002. *The frame-based module of the SUISEKI information extraction system*, *Intelligent Systems, IEEE,* 17; 2:14 - 20.

O. Bodenreider, T.C. Rindflesch, A. Burgun, 2002. *Unsupervised, corpus-based method for extending a biomedical terminology. Proc of the ACL-02 workshop on Natural language processing in the biomedical domain*: 53–60.

S. Brin 1998. *Extracting patterns and relations from the world wide web. WebDB Workshop at 6th International Conference on Extending Database Technology*

E. Buyko, S. Piao, Y. Tsuruoka, K. Tomanek , J.D. Kim, J. McNaught, U. Hahn, J. Su, and S. Ananiadou. 2007, *Bootstrep annotation scheme: Encoding information for text mining*, *Proc of the 4th Corpus Linguistics Conference*, Birmingham, July 27-30.

J. G. Caprosaso, W.A. Baumgartner, D.A. Randolph, K.B. Cohen, L. Hunter 2007. *Rapid pattern development for concept recognition systems: application to point mutations.*, *Journal of Bioinformatics and Computational Biology*, Vol. 5, No. 6, 12331259.

A. Cohen 2005. *Unsupervised gene/protein named entity normalization using automatically extracted dictionaries*. *Proc of the ACL-ISMB Workshop on Linking Biological Literature, Ontologies and Databases*: 17-24.

M. Collins, Y. Singer 1999. *Unsupervised Models for Named Entity Classification*. *EMNLP*

J.R. Curran, T Murphy, B Scholz 2007. *Minimizing Semantic Drift With Mutual Exclusion Bootstrapping*, *Proc of the 10th Conference of PACL*:172-180.

D. Day, J. Aberdeen, L. Hirschman, R. Kozierok, P. Robinson, M. Vilain 1997, *Mixed-initiative development of language processing systems*. *Proc of the 5th ACL Conference on Applied Natural Language Processing*

H. Fang, K. Murphy, Y. Jin, J.S. Kim, P.S. White 2006. *Human Gene Name Normalization using Text Matching with Automatically Extracted Synonym Dictionaries*. *Proc of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06*: 4148.

C. Friedman, P. Kra, H. Yu, M. Krauthammer, A. Rzhetsky. 2001. *GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles*. *Bioinformatics*, ;17 Suppl 1:S74-82.

M.A. Hearst 1992. *Automatic acquisition of hyponyms from large text corpora*, *Proc of the 14th conference on computational linguistics*.

D. Klein D, CD. Manning 2003. *Accurate Unlexicalized Parsing*, *Proc of the 41st Meeting of the Association for Computational Linguistics*, 2003; 423-30.

M. Krauthammer G. Nenadic 2004. *Term identification in the biomedical literature.*, *J Biomed Inform*, Dec;37(6):512-26.

S. Mukherjea, L.V. Subramaniam, G. Chanda, S. Sankararaman, R. Kothari, V.S. Batra, D.N. Bhardwaj, B.Srivastava 2004. *Enhancing a biomedical information extraction system with dictionary mining and context disambiguation*, *IBM Journal of Research and Development*, 48(5-6): 693-702

W. Pratt, M. Yetisgen-Yildiz 2003 *A Study of Biomedical Concept Identification: MetaMap vs. People*, *Proc AMIA Symp*, 529-533.

R. Snow, D. Jurafsky, A. Ng 2005. *Learning syntactic patterns for automatic hypernym discovery*, *Proc of the 17th Conference on Advances in Neural Information Processing Systems* MIT Press.

M. Thelen, E. Riloff 2002. *A Bootstrapping Method for Learning Semantic Lexicons Using Extraction Pattern Contexts*, *Proc of EMNLP*.

K. Tomanek, J Wermter, U Hahn. 2007. *An Approach to Text Corpus Construction which Cuts Annotation Costs and Maintains Reusability of Annotated Data*, *Proc of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*:486-495.

Y. Tsuruoka, J. Tsujii 2003, *Boosting Precision and Recall of Dictionary-Based Protein Name Recognition*, *Proc of the ACL 2003 Workshop on NLP in Biomedicine*:41-8.

Y. Tsuruoka, J. Tsujii 2004, *Improving the performance of dictionary-based approaches in protein name recognition*, *J of Biomed Inf* 37, 6; December: 461-470.

R. Xu, K. Supekar, A. Morgan, A.Das, A. Garber 2008. *Unsupervised Method for Automatic Construction of a Disease Dictionary from a Large Free Text Collection*, *Proc AMIA Symp*.