The Semantics of a Definiendum Constrains both the Lexical Semantics and the Lexicosyntactic Patterns in the Definiens

Hong Yu

Department of Health Sciences
University of Wisconsin-Milwaukee
Milwaukee, WI 53201
Hong.Yu@uwm.edu

Ying Wei

Department of Biostatistics
Columbia University
New York, NY 10032
Ying.Wei@columbia.com

Abstract

Most current definitional question answering systems apply one-size-fits-all lexicosyntactic patterns to identify definitions. By analyzing a large set of online definitions, this study shows that the semantic types of definienda constrain both lexical semantics and lexicosyntactic patterns of the definientia. For example, "heart" has the semantic type *[Body Part, Organ, or Organ Component]* and its definition (e.g., "heart locates between the lungs") incorporates semantic-type-dependent lexicosyntactic patterns (e.g., "*TERM locates* …") and terms (e.g., "lung" has the same semantic type *[Body Part, Organ, or Organ Component]*). In contrast, "AIDS" has a different semantic type *[Disease or Syndrome]*; its definition (e.g., "An infectious disease caused by human immunodeficiency virus") consists of different lexicosyntactic patterns (e.g., "*…causes by…*") and terms (e.g., "infectious disease" has the semantic type *[Disease or Syndrome]*). The semantic types are defined in the widely used biomedical knowledge resource, the Unified Medical Language System (UMLS).

1 Introduction

Definitional questions (e.g., "What is X?") constitute an important question type and have been a part of the evaluation at the Text Retrieval Conference (TREC) Question Answering Track since 2003. Most systems apply one-size-fits-all lexicosyntactic patterns to identify definitions (Liang et al. 2001; Blair-Goldensohn et al. 2004; Hildebrandt et al. 2004; Cui et al. 2005). For example, the pattern "*NP, (such as|like|including) query term*" can be used to identify the definition *"New research in mice suggests that drugs such as Ritalin quiet hyperactivity*" (Liang et al. 2001).

Few existing systems, however, have explored the relations between the semantic type (denoted as S_{DT}) of a definiendum (i.e., a defined term (DT)) and the semantic types (denoted as S_{Def}) of terms in its definiens (i.e., definition). Additionally, few existing systems have examined whether the lexicosyntactic patterns of definitions correlate with the semantic types of the defined terms.

By analyzing a large set of online definitions, this study shows that 1) S_{Def} correlates with S_{DT}, and 2) S_{DT} constrains the lexicosyntactic patterns of the corresponding definitions. In the following, we will illustrate our findings with the following four definitions:

a. ***Heart***[Body Part, Organ, or Organ Component]: The hollow[Spatial Concept] muscular[Spatial Concept] organ[Body Part, Organ, or Organ Component,Tissue] located[Spatial Concept] behind[Spatial Concept] the sternum[Body Part, Organ, or Organ Component] and between the lungs[Body Part, Organ, or Organ Component].

b. ***Kidney***[Body Part, Organ, or Organ Component]: The kidneys are a pair of glandular organs[Body Part, Organ, or Organ Component] located[Spatial Concept] in the abdominal_cavities[Body Part, Organ, or Organ Component] of mammals[Mammal] and reptiles[Reptile].

c. ***Heart attack***[Disease or Syndrome]: also called myocardial_infarction[Disease or Syndrome]; damage[Functional Concept] to the heart_muscle[Tissue] due to insufficient

Proceedings of the BioNLP Workshop on Linking Natural Language Processing and Biology at HLT-NAACL 06, pages 1–8,
New York City, June 2006.

blood supply[Organ or Tissue Function] for an extended[Spatial Concept] time_period[Temporal Concept].

d. ***AIDS***[Disease or Syndrome]: An infectious_disease[Disease or Syndrome] caused[Functional Concept] by human_immunodeficiency_virus[Virus].

In the above four definitions, the superscripts in [brackets] are the semantic types (e.g., [Body Part, Organ, or Organ Component] and [Disease or Syndrome]) of the preceding terms. A multiword term links words with the underscore "_". For example, "heart" IS-A [Body Part, Organ, or Organ Component] and "heart_muscle" IS-A [Tissue]. The semantic types are defined in the *Semantic Network (SN)* of the Unified Medical Language System (UMLS), the largest biomedical knowledge resource. Details of the UMLS and SN will be described in Section 2. We applied MMTx (Aronson et al. 2004) to automatically map a string to the UMLS semantic types. MMTx will also be described in Section 2.

Simple analysis of the above four definitions shows that given a defined term (DT) with a semantic type S_{DT} (e.g., [Body Part, Organ, or Organ Component]), terms that appear in the definition tend to have the same or related semantic types (e.g., [Body Part, Organ, or Organ Component] and [Spatial Concept]). Such observations were first reported as "Aristotelian definitions" (Bodenreider and Burgun 2002) in the limited domain of anatomy. (Rindflesch and Fiszman 2003) reported that the hyponym related to the definiendum must be in an IS-A relation with the hypernym that is related to the definiens. However, neither work demonstrated statistical patterns on a large corpus as we report in this study. Additionally, none of the work explicitly suggested the use of patterns to support question answering.

In addition to statistical correlations among semantic types, the lexicosyntactic patterns of the definitions correlate with S_{DT}. For example, as shown by sentences a~d, when S_{DT} is [Body Part, Organ, or Organ Component], its lexicosyntactic patterns include "…*located*…". In contrast, when S_{DT} is [Disease or Syndrome], the patterns include "…*due to*…" and "… *caused by*…".

In this study, we empirically studied statistical correlations between S_{DT} and S_{Def} and between S_{DT} and the lexicosyntactic patterns in the definitions. Our study is a result of detailed statistical analysis of 36,535 defined terms and their 226,089 online definitions. We built our semantic constraint model based on the widely used biomedical knowledge resource, the UMLS. We also adapted a robust information extraction system to generate automatically a large number of lexicosyntactic patterns from definitions. In the following, we will first describe the UMLS and its semantic types. We will then describe our data collection and our methods for pattern generation.

2 Unified Medical Language System

The Unified Medical Language System (UMLS) is the largest biomedical knowledge source maintained by the National Library of Medicine. It provides standardized biomedical concept relations and synonyms (Humphreys et al. 1998). The UMLS has been widely used in many natural language processing tasks, including information retrieval (Eichmann et al. 1998), extraction (Rindflesch et al. 2000), and text summarization (Elhadad et al. 2004; Fiszman et al. 2004).

The UMLS includes the Metathesaurus (MT), which contains over one million biomedical concepts and the Semantic Network (SN), which represents a high-level abstraction from the UMLS Metathesaurus. The SN consists of 134 semantic types with 54 types of semantic relations (e.g., *is-a* or *part-of*) that relate the semantic types to each other. The UMLS Semantic Network provides broad and general world knowledge that is related to human health. Each UMLS concept is assigned one or more semantic types.

The National Library of Medicine also makes available MMTx, a programming implementation of MetaMap (Aronson 2001), which maps free text to the UMLS concepts and associated semantic types. MMTx first parses text into sentences, then chunks the sentences into noun phrases. Each noun phrase is then mapped to a set of possible UMLS concepts, taking into account spelling and morphological variations; each concept is weighted, with the highest weight representing the most likely mapped concept. One recent study has evaluated MMTx to have 79% (Yu and Sable 2005) accuracy for mapping a term to the semantic

type(s) in a small set of medical questions. Another study (Lacson and Barzilay 2005) measured MMTx to have a recall of 74.3% for capturing the semantic types in another set of medical texts.

In this study, we applied MMTx to identify the semantic types of terms that appear in their definitions. For each candidate term, MMTx ranks a list of UMLS concepts with confidence. In this study, we selected the UMLS concept that was assigned with the highest confidence by MMTx. The UMLS concepts were then used to obtain the corresponding semantic types.

3 Data Collection

We collected a large number of online definitions for the purpose of our study. Specifically, we applied more than 1 million of the UMLS concepts as candidate definitional terms, and searched for the definitions from the World Wide Web using the *Google:Definition* service; this resulted in the downloads of a total of 226,089 definitions that corresponded to a total of 36,535 UMLS concepts (or 3.7% of the total of 1 million UMLS concepts). We removed from definitions the defined terms; this step is necessary for our statistical studies, which we will explain later in the following sections. We applied MMTx to obtain the corresponding semantic types.

4 Statistically Correlated Semantic Types

We then identified statistically correlated semantic types between S_{DT} and S_{Def} based on bivariate tabular chi-square (Fleiss 1981).

	Number of definitions that have STY_i	Number of definitions that don't have STY_i	Total
S_{Def}	$O(Def_i)$	$O(\underline{Def_i})$	N_{Def}
S_{All}	$O(All_i)$	$O(\underline{All_i})$	N_{All}
Total	N_i	$\underline{N_i}$	N

Specifically, given a semantic type $STY_{i,\ i=1,2,3,\ldots,\ 134}$ of any defined term, the observed numbers of definitions that were and were not assigned the STY_i are $O(Def_i)$ and $O(\underline{Def_i})$. *All* indicates the total 226,089 definitions. The observed numbers of definitions in which the semantic type STY_i did and did not appear were $O(All_i)$ and $O(\underline{All_i})$. 134 represents the total number of the UMLS semantic types. We applied formulas (1) and (2) to calculate expected frequencies and then the chi-square value (the degree of freedom is one). A high chi-square value indicates the importance of the semantic type that appears in the definition. We removed the defined terms from their definitions prior to the semantic-type statistical analysis in order to remove the bias introduced by the defined terms (i.e., defined terms frequently appear in the definitions).

$$E(Def_i) = \frac{N_{Def} * N_i}{N}, \quad E(\underline{Def_i}) = \frac{N_{Def} * \underline{N_i}}{N},$$
$$E(All_i) = \frac{N_{All} * N_i}{N}, \quad E(\underline{All_i}) = \frac{N_{All} * \underline{N_i}}{N} \qquad (1)$$

$$\chi^2 = \sum \frac{(E-O)^2}{E} \qquad (2)$$

To determine whether the chi-square value is large enough for statistical significance, we calculated its p-value. Typically, 0.05 is the cutoff of significance, i.e. significance is accepted if the corresponding p-value is less than 0.05. This criterion ensures the chance of false significance (incorrectly detected due to chance) is 0.05 for a single S_{DT}-S_{Def} pair. However, since there are 134*134 possible S_{DT}-S_{Def} pairs, the chance for obtaining at least one false significance could be very high. To have a more conservative inference, we employed a Bonferroni-type correction procedure (Hochberg 1988).

Specifically, let $p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$ be the ordered raw p-values, where *m* is the total number of S_{DT}-S_{Def} pairs. A S_{Def} is significantly associated with a S_{DT} if S_{Def}'s corresponding p-value $\le p_{(i)} \le \alpha/(m-i+1)$ for some i. This correction procedure allows the probability of at-least-one-false-significance out of the total m pairs is less than alpha (=0.05).

The number of definitions for each S_{DT} ranges from 4 ([Entity]), 10 ([Event]), 17 ([Vertebrate]) to 8,380 ([Amino Acid, Peptide, or Protein]) and 18,461 ([Organic Chemical]) in our data collection. As the power of a statistical test relies on the sample size, some correlated semantic types might be undetected when the number of available definitions is small. It is therefore worthwhile to know what the necessary sample size is in order to have a decent chance of detecting difference statistically.

For this task, we assume P_0 and P_1 are true probabilities that a STY will appear in N_{Def} and N_{All}. Based upon that, we calculated the minimal required number of sentences *n* such that the probability of statistical significance will be larger than or equal to 0.8. This sample size is determined based on the following two assumptions: 1) the observed frequencies are approximately normally distributed, and 2) we use chi-square significance to test the hypothesis $P_0 = P_1$ at significance level 0.05 ($\overline{P} = \frac{P_0 + P_1}{2}$).

$$n > \frac{(z_{0.025}\sqrt{2\overline{P}(1-\overline{P})} + z_{0.2}\sqrt{P_1(1-P_1)+P_0(1-P_0)})^2}{(P_0 - P_1)^2} \qquad (3)$$

5 Semantic Type Distribution

Our null hypothesis is that given any pair of $\{S_{DT}(X), S_{DT}(Y)\}$, $X \neq Y$, where X and Y represent two different semantic types of the total 134 semantic types, there are no statistical differences in the distributions of the semantic types of the terms that appear in the definitions.

We applied the bivariate tabular chi-square test to measure the semantic type distribution. Following similar notations to Section 4, we use $\underline{O}_{Xi}$ and $\underline{O}_{Yi}$ for the corresponding frequencies of not being observed in $S_{Def}(X)$ and $S_{Def}(Y)$.

For each semantic type STY, we calculate the expected frequencies of being observed and not being observed in $S_{Def}(X)$ and $S_{Def}(Y)$, respectively, and their corresponding chi-square value according to formulas (3) and (4):

$$E_{X_i} = \frac{N_{X_i} * (O_{X_i} + O_{Y_i})}{N_{X_i} + N_{Y_i}},\ \underline{E}_{X_i} = \frac{N_{X_i} * (\underline{O}_{X_i} + \underline{O}_{Y_i})}{N_{X_i} + N_{Y_i}},$$

$$E_{Y_i} = \frac{N_{Y_i} * (O_{X_i} + O_{Y_i})}{N_{X_i} + N_{Y_i}}, \underline{E}_{Y_i} = \frac{N_{Y_i} * (\underline{O}_{X_i} + \underline{O}_{Y_i})}{N_{X_i} + N_{Y_i}} \qquad (4)$$

$$\chi^2_{X,Y,i} = \sum \frac{(E_{X_i} - O_{X_i})^2}{E_{X_i}} + \sum \frac{(E_{Y_i} - O_{Y_i})^2}{E_{Y_i}} \qquad (5)$$

where N_X and N_Y are the numbers of sentences in $S_{Def}(X)$ and $S_{Def}(Y)$, respectively, and in both (4) and (5), $i = 1,2,...,134$, and $(X, Y) = {}_{1,2,...,\ 134}$ and $X \neq Y$. The degree of freedom is *1*. The chi-square value measures whether the occurrences of STY_i, are equivalent between $S_{Def}(X)$ and $S_{Def}(Y)$. The same multiple testing correction procedure will be used to determine the significance of the chi-square value. Note that if at least one STY_i has been detected to be statistically significant after multiple-testing correction, the distributions of the semantic types are different between $S_{Def}(X)$ and $S_{Def}(Y)$.

6 Automatically Identifying Semantic-Type-Dependent Lexicosyntactic Patterns

Most current definitional question answering systems generate lexicosyntactic patterns either manually or semi-automatically. In this study, we automatically generated large sets of lexicosyntactic patterns from our collection of online definitions. We applied the information extraction system Autoslog-TS (Riloff and Philips 2004) to automatically generate lexicosyntactic patterns in definitions. We then identified the statistical correlation between the semantic types of defined terms and their lexicosyntactic patterns in definitions.

AutoSlog-TS is an information extraction system that is built upon AutoSlog (Riloff 1996). AutoSlog-TS automatically identifies extraction patterns for noun phrases by learning from two sets of un-annotated texts *relevant* and *non-relevant*. AutoSlog-TS first generates every possible lexicosyntactic pattern to extract every noun phrase in both collections of text and then computes statistics based on how often each pattern appears in the relevant text versus the background and outputs a ranked list of extraction patterns coupled with statistics indicating how strongly each pattern is associated with *relevant* and *non-relevant* texts.

We grouped definitions based on the semantic types of the defined terms. For each semantic type, the *relevant* text incorporated the definitions, and the *non-relevant* text incorporated an equal number of sentences that were randomly selected from the MEDLINE collection. For each semantic type, we applied AutoSlog-TS to its associated *relevant* and *non-relevant* sentence collections to generate lexicosyntactic patterns; this resulted in a total of 134 sets of lexicosyntactic patterns that corresponded to different semantic types of defined terms. Additionally, we identified the common lexicosyntactic patterns across the semantic types and ranked the lexicosyntactic patterns based on their frequencies across semantic types.

We also identified statistical correlations between S_{DT} and the lexicosyntactic patterns in definitions based on chi-square statistics that we have described in the previous two sections. For formula 1~4, we replaced each STY with a lexicosyntactic pattern. Our null hypothesis is that given any S_{DT}, there are no statistical differences in the distributions of the lexicosyntactic patterns that appear in the definitions.

[Body Part, Organ, or Organ Component]:
[Body Part, Organ, or Organ Component]
[Spatial Concept]
[Tissue]
[Body Location or Region]
[Medical Device]

[Disease or Syndrome]:
[Disease or Syndrome]
[Pathologic Function]
[Body Part, Organ, or Organ Component]
[Sign or Symptom]
[Finding]

[Organization]:
[Organization]
[Regulation or Law]
[Governmental or Regulatory Activity]
[Social Behavior]
[Occupational Activity]

Figure 1: A list of semantic types of defined terms with the top five statistically correlated semantic types (P<<0.0001) that appear in their definitions.

7 Results

Our chi-square statistics show that for any pair of semantic types $\{S_{DT}(X), S_{DT}(Y)\}$, $X \neq Y$, the distributions of S_{Def} are statistically different at alpha=0.05; the results show that the semantic types of the defined terms correlate to the semantic types in the definitions. Our results also show that the syntactic patterns are distributed differently among different semantic types of the defined terms (alpha=0.05).

Our results show that many semantic types that appear in definitions are statistically correlated with the semantic types of the defined terms. The average number and standard deviation of statistically correlated semantic types is 80.6±35.4 at P<<0.0001.

Figure 1 shows three S_{DT} ([Body Part, Organ, or Organ Component], [Disease or Syndrome], and [Organization]) with the corresponding top five statistically correlated semantic types that appear in their definitions. Our results show that in a total of 112 (or 83.6%) cases, S_{DT} appears as one of the top five statistically correlated semantic types in S_{Def}, and that in a total of 94 (or 70.1%) cases, S_{DT} appears at the top in S_{Def}. Our results indicate that if a definitional term has a semantic type S_{DT}, then the terms in its definition tend to have the same or related semantic types.

We examined the cases in which the semantic types of definitional terms do not appear in the top five semantic types in the definitions. We found that in all of those cases, the total numbers of definitions that were used for statistical analysis were too small to obtain statistical significance. For example, when S_{DT} is "Entity", the minimum size for a S_{Def} was 4.75, which is larger than the total number of the definitions (i.e., 4). As a result, some actually correlated semantic types might be undetected due to insufficient sample size.

Our results also show that the lexicosyntactic patterns of definitional sentences are S_{DT}-dependent. Our results show that many lexicosyntactic patterns that appear in definitions are statistically correlated with the semantic types of defined terms. The average number and standard deviation of statistically correlated lexico-syntactic patterns is 1656.7±1818.9 at P<<0.0001. We found that the more definitions an S_{DT} has, the more lexicosyntactic patterns.

Figure 2 shows the top 10 lexicosyntactic patterns (based on chi-square statistics) that were captured by Autoslog-TS with three different S_{DT}; namely, *[Disease or Syndrome], [Body Part, Organ, or Organ Component],* and *[Organization]*. Figure 3 shows the top 10 lexicosyntactic patterns ranked by AutoSlog-TS which incorporated the frequencies of the patterns (Riloff and Philips 2004).

Figure 4 lists the top 30 common patterns across all different semantic types S_{DT}. We found that many common lexicosyntactic patterns (e.g., "…known as…", "…called", "…include…") have been identified by other research groups through either manual or semi-automatic pattern discovery (Blair-Goldensohn et al. 2004).

[Disease or Syndrome]	[Body Part, Organ, or Organ Component]	[Organization]
Np_Prep_<NP>__INFLAMMATION_OF	Np_Prep_<NP>__PART_OF	Np_Prep_<NP>__UNION_IN
ActVp_Prep_<NP>__CHARACTERIZED_BY	<subj>_ActVp__LOCATED	<subj>_AuxVp_Dobj__BE_COURT
<subj>_ActVp__CHARACTERIZED	ActVp_<dobj>__CALLED	Np_Prep_<NP>__ORGANIZATION_TO
ActVp_<dobj>__CALLED	Np_Prep_<NP>__PORTION_OF	Np_Prep_<NP>__GOVERNMENT_WITH
<subj>_ActVp__OCCURS	<subj>_ActVp__CALLED	Subj_AuxVp_<dobj>__BE_ARMY
Np_Prep_<NP>__LOSS_OF	Np_Prep_<NP>__SIDE_OF	Np_Prep_<NP>__WORSHIP_FOR
<subj>_ActVp__CAUSES	ActVp_<dobj>__CONTAINS	ActVp_Prep_<NP>__FORMED_FOR
<subj>_ActVp__INCLUDE	Np_Prep_<NP>__BASE_OF	<subj>_ActVp_Dobj__OWN_PROPERTY
ActVp_<dobj>__CAUSES	Np_Prep_<NP>__ORGAN_IN	Subj_AuxVp_<dobj>__HAVE_CORPORATION
Np_Prep_<NP>__FORM_OF	Np_Prep_<NP>__LAYER_OF	Subj_AuxVp_<dobj>__BE_ORGANIZATIONS

Figure 2: The top 10 lexicosyntactic patterns that appear in definitions based on chi-square statistics. The defined terms have one of the three semantic types *[Disease_or_Syndrome], [Body Part, Organ, or Organ Component],* and *[Organization]*.

[Disease or Syndrome]	[Body Part, Organ, or Organ Component]	[Organization]
Np_Prep_<NP>__INFLAMMATION_OF	Np_Prep_<NP>__PART_OF	Np_Prep_<NP>__GROUP_OF
<subj>_ActVp__DISEASE	<subj>_ActVp__LOCATED	Np_Prep_<NP>__HOUSE_OF
ActVp_Prep_<NP>__CHARACTERIZED_BY	ActVp_<dobj>__CALLED	Np_Prep_<NP>__PLACE_OF
<subj>_ActVp__CHARACTERIZED	Np_Prep_<NP>__PORTION_OF	ActVp_<dobj>__INCLUDING
ActVp_<dobj>__CALLED	<subj>_ActVp__CALLED	<subj>_ActVp__ESTABLISHED
ActVp_Prep_<NP>__CAUSED_BY	Np_Prep_<NP>__SIDE_OF	<subj>_ActVp__FORMED
<subj>_ActVp__OCCURS	ActVp_<dobj>__CONTAINS	Np_Prep_<NP>__COURT_OF
Np_Prep_<NP>__LOSS_OF	Np_Prep_<NP>__BASE_OF	<subj>_ActVp__INCLUDE
<subj>_ActVp__CAUSED	Np_Prep_<NP>__ORGAN_IN	ActVp_<dobj>__SEE
<subj>_ActVp__CAUSES	Np_Prep_<NP>__LAYER_OF	ActVp_Prep_<NP>__REFERS_TO

Figure 3: The top 10 lexicosyntactic patterns ranked by Autoslog-TS. The defined terms have one of the three semantic types *[Disease_or_Syndrome], [Body Part, Organ, or Organ Component],* and *[Organization]*.

1. ActVp_<dobj>__SEE	11. <subj>_PassVp__USED	21. <subj>_ActVp__INCLUDES
2. <subj>_ActVp__USED	12. ActVp_Prep_<NP>__KNOWN_AS	22. <subj>_ActVp__INCLUDING
3. ActVp_<dobj>__CALLED	13. ActVp_<dobj>__INCLUDES	23. <subj>_ActVp__LIKE
4. Subj_AuxVp_<dobj>__BE_IT	14. Np_Prep_<NP>__TYPE_OF	24. <subj>_ActVp__DISEASE
5. <subj>_ActVp__CALLED	15. ActVp_Prep_<NP>__USED_IN	25. <subj>_ActVp__REFERS
6. Np_Prep_<NP>__PART_OF	16. ActVp_<dobj>__LIKE	26. Np_Prep_<NP>__NUMBER_OF
7. ActVp_<dobj>__INCLUDING	17. Np_Prep_<NP>__ONE_OF	27. ActVp_Prep_<NP>__FOUND_IN
8. <subj>_ActVp__INCLUDE	18. Np_Prep_<NP>__FORM_OF	28. <subj>_ActVp__KNOWN
9. ActVp_<dobj>__INCLUDE	19. Np_Prep_<NP>__GROUP_OF	29. Np_Prep_<NP>__PROCESS_OF
10. ActVp_Prep_<NP>__REFERS_TO	20. <subj>_PassVp__CALLED	30. <subj>_ActVp__OCCURS

Figure 4: The top 30 common lexicosyntactic patterns generated across patterns with different S_{DT} .

8 Discussion

The statistical correlations between S_{DT} and S_{Def} may be useful to enhance the performance of a definition-question-answering system by at least two means. First, the semantic types may be useful for word sense disambiguation. A simple application is to rank definitional sentences based on the distributions of the semantic types of terms in the definitions to capture the definition of a specific sense. For example, a biomedical definitional question answering system may exclude the definition of other senses (e.g., "feeling" as shown in the sentence "The locus of feelings and intuitions; 'in your heart you know it is true'; 'her story would melt your heart.'") if the semantic types that define "heart" do not include [Body Part, Organ, or Organ Component] of terms other than "heart".

Secondly, the semantic-type correlations may be used as features to exclude non-definitional sentences. For example, a biomedical definitional question answering system may exclude the following non-definitional sentence "Heart rate was

unaffected by the drug" because the semantic types in the sentence do not include [Body Part, Organ, or Organ Component] of terms other than "heart".

S_{DT}-dependent lexicosyntactic patterns may enhance both the recall and precision of a definitional question answering system. First, the large sets of lexicosyntactic patterns we generated automatically may expand the smaller sets of lexicosyntactic patterns that have been reported by the existing question answering systems. Secondly, S_{DT}-dependent lexicosyntactic patterns may be used to capture definitions.

The common lexicosyntactic patterns we identified (in Figure 4) may be useful for a generic definitional question answering system. For example, a definitional question answering system may implement the most common patterns to detect any generic definitions; specific patterns may be implemented to detect definitions with specific S_{DT}.

One limitation of our work is that the lexicosyntactic patterns generated by Autoslog-TS are within clauses. This is a disadvantage because 1) lexicosyntactic patterns can extend beyond clauses (Cui et al. 2005) and 2) frequently a definition has multiple lexicosyntactic patterns. Many of the patterns might not be generalizible. For example, as shown in Figure 2, some of the top ranked patterns (e.g., "Subj_AuxVp_<dobj>_BE_ARMY>") identified by AutoSlog-TS may be too specific to the text collection. The pattern-ranking method introduced by AutoSlog-TS takes into consideration the frequency of a pattern and therefore is a better ranking method than the chi-square ranking (shown in Figure 3).

9 Related Work

Systems have used named entities (e.g., "PEOPLE" and "LOCATION") to assist in information extraction (Agichtein and Gravano 2000) and question answering (Moldovan et al. 2002; Filatova and Prager 2005). Semantic constraints were first explored by (Bodenreider and Burgun 2002; Rindflesch and Fiszman 2003) who observed that the principle nouns in definientia are frequently semantically related (e.g., hyponyms, hypernyms, siblings, and synonyms) to definiena. Semantic constraints have been introduced to definitional question answering (Prager et al. 2000; Liang et al. 2001). For example, an artist's work must be completed between his birth and death (Prager et al. 2000); and the hyponyms of defined terms might be incorporated in the definitions (Liang et al. 2001). Semantic correlations have been explored in other areas of NLP. For example, researchers (Turney 2002; Yu and Hatzivassiloglou 2003) have identified semantic correlation between words and views: positive words tend to appear more frequently in positive movie and product reviews and newswire article sentences that have a positive semantic orientation and vice versa for negative reviews or sentences with a negative semantic orientation.

10 Conclusions and Future Work

This is the first study in definitional question answering that concludes that the semantics of a definiendum constrain both the lexical semantics and the lexicosyntactic patterns in the definition. Our discoveries may be useful for the building of a biomedical definitional question answering system.

Although our discoveries (i.e., that the semantic types of the definitional terms determine both the lexicosyntactic patterns and the semantic types in the definitions) were evaluated with the knowledge framework from the biomedical, domain-specific knowledge resource the UMLS, the principles may be generalizable to any type of semantic classification of definitions. The semantic constraints may enhance both recall and precision of one-size-fits-all question answering systems, which may be evaluated in future work.

As stated in the Discussion session, one disadvantage of this study is that the lexicosyntactic patterns generated by Autoslog-TS are within clauses. Future work needs to develop pattern-recognition systems that are capable of detecting patterns across clauses.

In addition, future work needs to move beyond lexicosyntactic patterns to extract semantic-lexicosyntactic patterns and to evaluate how the semantic-lexicosyntactic patterns can enhance definitional question answering.

Acknowledgement: The author thanks Sasha Blair-Goldensohn, Vijay Shanker, and especially the three anonymous reviewers who provide valuable critics and comments. The concepts “Definiendum” and “Definiens” come from one of the reviewers’ recommendation.

References

Agichtein E, Gravano L (2000) Snowball: extracting relations from large plain-text collections. . Paper presented at Proceedings of the 5th ACM International Conference on Digital Libraries

Aronson A (2001) Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. Paper presented at American Medical Information Association

Aronson A, Mork J, Gay G, Humphrey S, Rogers W (2004) The NLM Indexing Initiative's Medical Text Indexer. Paper presented at MedInfo 2004

Blair-Goldensohn S, McKeown K, Schlaikjer A (2004) Answering Definitional Questions: A Hybrid Approach. In: Maybury M (ed) New Directions In Question Answering. AAAI Press

Bodenreider O, Burgun A (2002) Characterizing the definitions of anatomical concepts in WordNet and specialized sources. Paper presented at The First Global WordNet Conference

Cui H, Kan M, Cua T (2005) Generic soft pattern models for definitional question answering. . Paper presented at The 28th Annual International ACM SIGIR Salvado, Brazil

Eichmann D, Ruiz M, Srinivasan P (1998) Cross-language information retrieval with the UMLS metathesaurus. Paper presented at SIGIR

Elhadad N, Kan M, Klavans J, McKeown K (2004) Customization in a unified framework for summarizing medical literature. Journal of Artificial Intelligence in Medicine

Filatova E, Prager J (2005) Tell me what you do and I'll tell you what you are: learning occupation-related activities for biographies. Paper presented at HLT/EMNLP 2005. Vancouver, Canada

Fiszman M, Rindflesch T, Kilicoglu H (2004) Abstraction Summarization for Managing the Biomedical Research Literature. Paper presented at HLT-NAACL 2004: Computational Lexical Semantic Workshop

Fleiss J (1981) Statistical methods for rates and proportions.

Hildebrandt W, Katz B, Lin J (2004) Answering definition questions with multiple knowledge sources. . Paper presented at HLT/NAACL

Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75:800-802

Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO (1998) The Unified Medical Language System: an informatics research collaboration. J Am Med Inform Assoc 5:1-11.

Lacson R, Barzilay R (2005) Automatic processing of spoken dialogue in the hemodialysis domain. Paper presented at Proc AMIA Symp

Liang L, Liu C, Xu Y-Q, Guo B, Shum H-Y (2001) Real-time texture synthesis by patch-based sampling. ACM Trans Graph 20:127--150

Moldovan D, Harabagiu S, Girju R, Morarescu P, Lacatusu F, Novischi A, Badulescu A, Bolohan O (2002) LCC tools for question answering. Paper presented at The Eleventh Text REtrieval Conference (TREC 2002)

Prager J, Brown E, Coden A, Radev D (2000) Quesiton-answering by predictive annotation. Paper presented at Proceeding 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval

Riloff E (1996) Automatically generating extraction patterns from untagged text. . Paper presented at AAAI-96

Riloff E, Philips W (2004) An introduction to the Sundance and AutoSlog Systems. Technical Report #UUCS-04-015. University of Utah School of Computing.

Rindflesch T, Tanabe L, Weinstein J, Hunter L (2000) EDGAR: extraction of drugs, genes and relations from the biomedical literature. Pac Symp Biocomput:517-528.

Rindflesch TC, Fiszman M (2003) The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text. J Biomed Inform 36:462-477

Turney P (2002) Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. Paper presented at ACL 2002

Yu H, Hatzivassiloglou V (2003) Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. Paper presented at Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP 2003)

Yu H, Sable C (2005) Being Erlang Shen: Identifying answerable questions. Paper presented at Nineteenth International Joint Conference on Artificial Intelligence on Knowledge and Reasoning for Answering Questions