Data Reduction for Continuum of Care: An Exploratory Study Using the Predicate-Argument Structure to Pre-process Radiology Sentences for Measurement of Semantic Similarity

Eric Newsom[1] and Josette F. Jones[2]

[1] Indiana University Health, Indianapolis, IN, USA
enewsom@iuhealth.org
[2] Indiana University-Purdue University Indianapolis, Indianapolis, IN, USA
jofjones@iupui.edu

Abstract. In the clinical setting, continuum of care depends on integrated information services to assure a smooth progression for patient centered care, and these integrated information services must understand past events and personal circumstances to make care relevant. Clinicians face a problem that the amount of information produced in disparate electronic clinical notes is increasing to levels incapable of being processed by humans. Clinicians need a function in information services that can reduce the free text data to a message useful at time of care. Information extraction (IE) is a sub-field of natural language processing with the goal of data reduction of unstructured free text. Pertinent to IE is an annotated corpus that frames how IE methods should create a logical expression necessary for processing meaning of text. This study explores and reports on the requirements to using the predicate-argument statement (PAS) as the framework. A convenient sample from a prior study with ten synsets of 100 unique sentences from radiology reports deemed by domain experts to mean the same thing will be the text from which PAS structures are formed. Through content analysis of pattern recognition, findings show PAS is a feasible framework to structure sentences for semantic similarity measurement.

Keywords: Information Extraction, Predicate-Argument Structure, Semantic Similarity.

1 Introduction

Today's knowledge worker has far too much published information to review for conducting his/her professional practice [1]. In healthcare, the majority of knowledge is in the form of free text such as professional journals and patient clinical notes [2]. These forms of free text have been linked to profits [3] and are central to a health system reliant on an integration of services to deliver patient centered care over a continuum [3,4]. Informational continuity forms from past events and personal circumstances [6] allowing for the clinician to establish a personal relationship and

C. Stephanidis and M. Antona (Eds.): UAHCI/HCII 2013, Part III, LNCS 8011, pp. 516–525, 2013.
© Springer-Verlag Berlin Heidelberg 2013

cooperation [7]. These clinical notes, therefore, provide the patient specific data necessary for continuum of care. The continuum of care can be greatly enhanced if means exist to synthesize from disparate clinical notes a meaningful message at the point in time for delivery of care. Analyzing these unique forms of knowledge by a machine-driven mechanism in order to decrease the knowledge worker's burden of reading is a goal of natural language processing (NLP). For data reduction to occur as a normal function in continuity of care, NLP methods use information extraction (IE) methods [2], which depend on an annotated corpus [1] to model structuring of unstructured text as a pre-process before semantic comparisons. To accomplish this, an annotation method has to be used that can retain the semantics of the entire sentence. One method uses a form of semantic role labeling (SRL) called predicate-argument structures (PAS) for IE of free text and appears to be the most commonly researched and used method for analysis of entire sentence structures [1,8-18].

For this study, PAS will serve as the annotation framework, but the measurement that will be used to determine the semantic equivalence for data reduction between two PAS elements will be semantic similarity. Semantic similarity is degree of closeness between two different texts and attempts to replicate how humans represent relationships within these varying vocabulary expressions to formulate meaning of experiences [19]. This study will evaluate the feasibility to use PAS as a pre-processing step for comparing semantic similarity of sentences.

1.1 Problem Statement

To test semantic similarity, concepts need to be represented at the atomic level in order to understand how broad or narrow one text element is to another [20]. Successful studies that have tested for semantic similarity have focused on terms or lexical elements [21-24]. These studies attest to the idea that less complexity of an expressed thought the easier it is to measure semantic similarity. In another study that comes closest to evaluating sentential semantic similarity through a method called Named Entity Recognition (NER), the method in [25] fails with issues of complex synonymy, overlapping text span, and structured interpretations. [25] attempts to look at the whole meaning of a sentence, but the study underlines the difficulty of comparing the equivalency between two sentences when dissected and chunked into smaller parts. In a PAS annotation study that closely resembles this one, annotation of semantics is not evaluated for semantic similarity [26]. It remains to be evaluated if PAS frames can rise to a level of complexity to analyze sentential semantic similarity.

2 Review of Literature

The contribution of PAS to IE is its ability to retain structure of sentences [11], contribute to inductive learning [16], and facilitate mapping of arguments to ontological references [15]. These are three components essential for successful NLP [3]. Typically, PAS development is guided by guidelines introduced through the Propbank project [18,26] whereby a usage of a predicate's sense is referenced by a RolesetID frame (see.01, be.02) in its knowledgebase. The RolesetID frame provides argument definitions to mark textual boundaries in sentence annotation.

While guidelines instruct a process of annotation of PAS, literature suggests the annotation process incorporate certain considerations. [15,16] contend that modification of a PAS frame is encouraged to correctly identify observations of a domain's structure necessary for semantic translation. For a corpus of radiology reports, a study by [27] frequently encountered missing predicates and implied concepts. These findings concerning a radiological corpus could mean that annotation will identify arguments and predicates not anchored in text. However, if domain experts make this kind of inference then the purpose of annotation serves to mimic the process of human judgment. For sentences lacking a predicate, this study will have to determine a method in the annotation process for measuring successfulness of constructing complete sentences from incomplete sentences and inferring implied/incomplete text.

If method of annotation is PAS, then focus is on the predicate, and to assure that predicates function to guide argument development, [12] argue that verbs should be classified based on syntactical structures present in corpus. For example, does the predicate accept a direct object or does it not. Such syntactical analysis of how a sentence is structured around specific verbs could help understand the arguments. Another issue that has to be addressed before annotating PAS frames is to understand how the corpus handles nominalization of verbs [13]. Nominalized verbs take different forms, such as gerunds, and these forms may have arguments despite not being the predicate of the sentence. This study will have to make a conclusion if the Propbank annotation guidelines and knowledge base can serve as the formal schema to build a PAS frame from a corpus that has unique usages of verbs and how those usages have to be annotated. The complexity of using PAS as a formal schema to measure sentential semantic similarity lies not in forming syntactical phrases but in examining similarity of text phrases of one sentence to text phrases of another. This kind of measurement goes beyond the discreteness of word to word comparison and requires a representational scheme to assist with matching which phrase of one sentence should be compared to a phrase of another sentence. To adhere to principles of semantic similarity, this representational scheme not only has to address syntax and conceptual needs in a sentence, but it has to provide a foundation to capture the accumulative sense formed by concepts in the sentence [15]. The goal, then, is to have a representational scheme whereby the output of syntactical phrases formed from one sentence are appropriately matched and compared to those of another sentence, and this representational scheme will serve as the basis for pattern discovery of synonymy [28].

Creating this scheme could pose the biggest obstacle to using the PAS. [27] state PAS cannot scale; however, it is not known if this statement is true or not. The problem of scalability of PAS is presented when two semantically equivalent sentences have two different predicates. In such cases, it is possible that content of the arguments and modifiers of one sentence may not match to the other predicate's corresponding arguments and modifiers. In some cases, there may be no text to compare. Table 1 shows how three semantically equivalent sentences present the problem of scalability with three difference predicate senses.

Table 1. --Scalability Problem of PAS

Sentence 1	The lungs demonstrate left basilar atelectasis or infiltrate.		
Sentence 2	There is interval clearing of the left lower lobe infiltrate.		
Sentence 3	The left lower lobe infiltrate is identified.		
	Sentence 1	Sentence 2	Sentence 3
RolesetID	demonstrate.01	be.02	identify.01
Predicate Use	*'show off'*	*'existential'*	*'label, call'*
Arg0	The lungs		
Arg1	**atelectasis or infiltrate**	interval clearing	**infiltrate**
Arg2			
ArgM-Locative	left basilar	of the left lower lobe	The left lower lobe
ArgM-Cause		**infiltrate**	

In each sentence, PAS annotates the complete syntactical structure of the sentence; therefore, the complete sense of each sentence is maintained. But when the goal is to compare the semantics across each role, problems arise from the fact that each predicate assumes a different structure for role assignments. It is evident that when trying to measure semantic similarity of varying predicates, not every role will contain like content. Arg1 of sentence 1 & 3 have semantic similarity but Arg1 of sentence 2 is not semanticly similar despite the fact that the defined roles of the verb places the actual content that could be compared in the modifier 'cause'. The question then is how is it possible to 'cross' compare arguments for a synset comprised of various predicates? Is it possible to compare differing semantic roles? This study will investigate patterns among synsets with varying predicates for feasibility of an algorithm allowing cross comparing of semantic roles.

3 Methods, Design, and Sampling

This study used a traditional annotation schema for clinical corpus [29]. Guidelines for the schema were adopted from the Propbank annotation [30] and used the Unified Verb Index [31]to facilitate construction of PAS frames. The reader is directed to [30] for further understanding of PAS. While the study is primarily an exploratory, non-experimental, methodological study, it will use a content analysis process for pattern recognition of meaning in text developed for NLP [32].

A convenience sample from a prior study [27] is used as the data set. The data set represents ten synsets. Each synset has 100 randomly selected sentences (n=1000) from original synsets containing more than 1,000 sentences of an ongoing annotation project. A propositional sentence serves as the representing sentence for the synset (see Table 2).

Table 2. --Proposition Sentences of Synsets

Synset	Propositional Sentence of Synset
1.	The endotracheal tube is above the carina.
2.	There is no pneumothorax.
3.	There is a left lower lobe pulmonary infiltrate(s).
4.	The pulmonary vessels are prominent.
5.	A posterior anterior (PA) chest x-ray was performed.
6.	The gray white matter differentiation of the brain is normal.
7.	The intervertebral disc heights are normal.
8.	There are pelvic phlebolith(s).
9.	There is small vessel ischemic disease of the brain.
10.	The lungs are diffusely hazy bilaterally.

3.1 Data Analysis/Content Analysis

One researcher validated and analyzed the data but used alternative means reported in [33] to address validity. The content analysis method was adapted from [15] to address the dominant problem of scalability. In that study, development of PAS begins by identifying the NP and VP. In this study, the NP and VP was annotated (see Figure 1) for each of the proposition sentences of the 10 synsets and extended to annotate respective modifiers.

Synset Proposition: The endotracheal tube is above the carina.
NP VP

Fig. 1. Synset 1 in Table 2

The intent is to establish a baseline for possible pattern recognition of semantically equivalent phrases not annotated to identical arguments in PAS frames due to varying uses of predicates.

4 Findings

For sentences with missing predicates, a process was applied whereby the annotator transformed the incomplete sentence to a complete, firstly, with the predicate RolesetID 'be.02' ('there are'/'there is'<NP>). If the sentence did not make sense with 'be.02' usage, 'be.01' RolesetID was applied (<NP> 'is'/'are'<VP>). If that transformed structure did not make sense, the annotator used any predicate that made the sentence understandable. 260 sentences were missing a predicate and the results are reported in Figure 2. The summary indicates that a radiology sentence missing a predicate can more than likely be completed with 'be.02'.

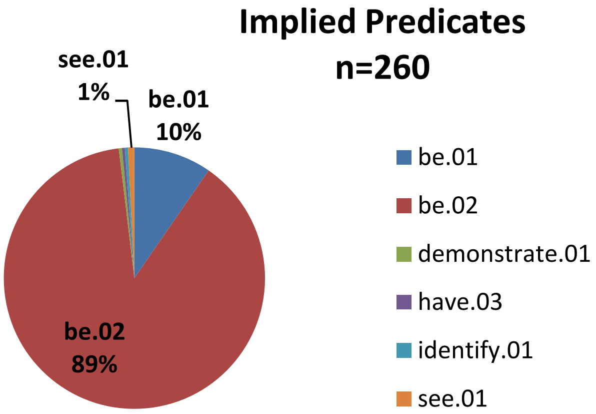

Fig. 2. Summary of All Synsets for Sentences Missing A Predicate

Figure 3 shows a frequency summary of predicate usage across sysnsets that clearly shows the dominate usage of the predicates 'to be', be.01 and be.02. With over 2/3 of the predicates expressed in the dataset using an existential form of 'to be', it is evident that Arg0 is absent. In fact, with non-'to be' predicates, less than 1% annotated text to Arg0. Arg0 represents the role of an agent or someone doing the action. In contrast, all sentences annotated with Arg1 (entity that receives action). According to [30], this indicates that predicates in this dataset are 'externally caused'. With externally caused predicates, the explanation for the motive or stimulate in the sentence is implied. The implication that predicates in a corpus are externally-caused vs. internally-caused suggests that the writer has a great deal of leniency to express a thought. There is an element that use of language in such a corpus is generalizable. A radiology report merely has to present observations and interpretation of those observations. It does not have to explain the actions of the disease process requiring use of internally-cause predicates. Because of the generalizability of text annotated to an argument for externally-caused predicates, it is possible to compare the content of an argument with a 'defined' semantic role label to content of an argument of a differently 'defined' semantic role label.

If comparison of differing semantic role labels is feasible in a radiology corpus, then a predictable pattern is necessary to establish an algorithm for IE processing. For each of the sysnets, a flowchart was developed to indicate based on the RolesetID, where the content of NP, VP, and modifiers of proposition sentence is annotated in candidate sentences. An example of the flow chart is Figure 4. Across all synsets, it is estimated that the uncovered pattern will be an algorithm with a content coverage of no less than 96%. Based on a predicate's usage (RolesetID), the necessary content of

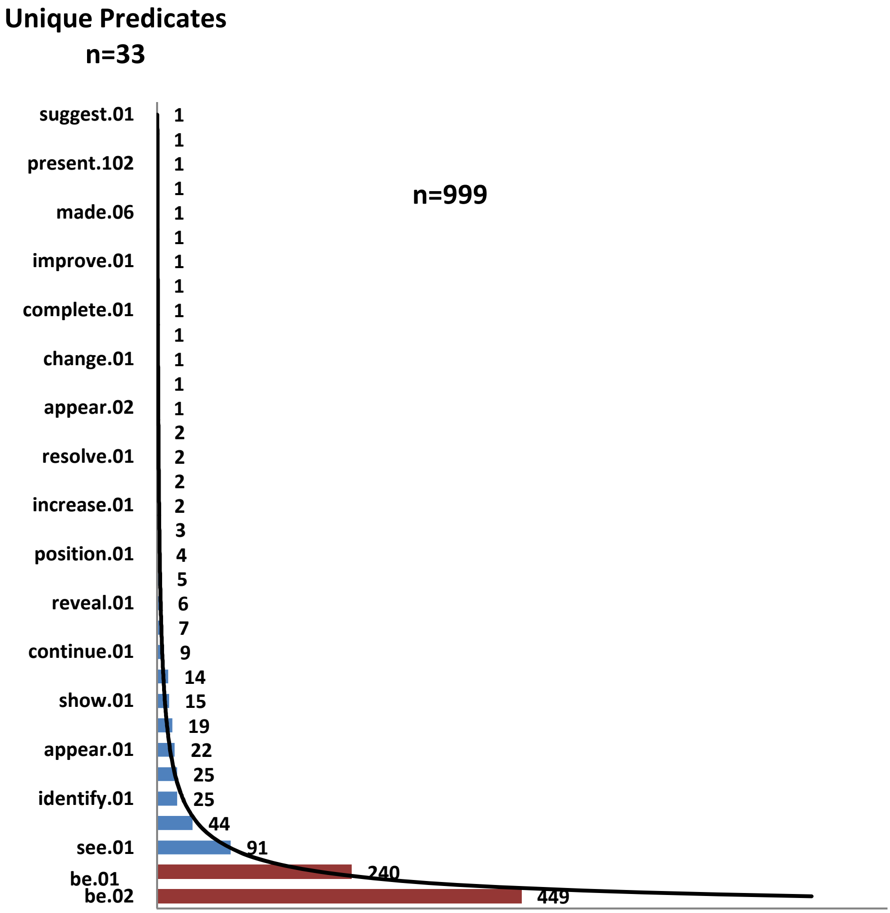

Fig. 3. RolesetID Frequencies in Dataset Corpus

arguments/modifiers of the candidate sentence needed to compare to those of the proposition sentence can be predicted. For RolesetIDs ‘be.01’ and ‘be.02’, conflicting patterns were often found; however, in all cases, patterns of usage determine what circumstances apply to one usage sense vs. another. In Figure 4, the VP concept of proposition sentence is found in Arg2 as long as Arg2 does not have the terms ‘place’, ‘position’, and ‘present’. Otherwise, for RolesetIDs ‘be.01’ and ‘be.02’, the VP concept is found in the PAS modifier ‘Locative’ of the candidate sentence.

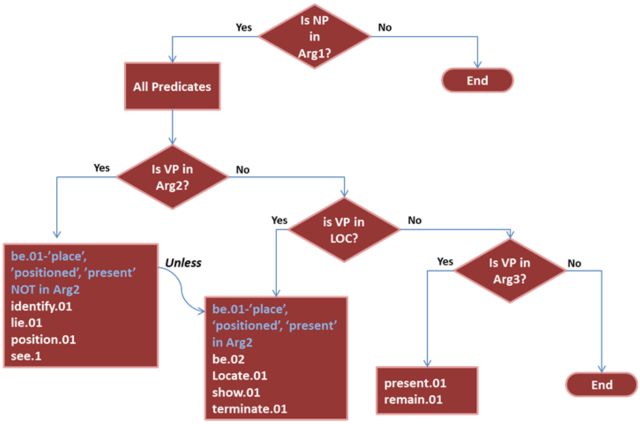

Fig. 4. --Algorithm Predicting Location of Equivelent Content of Proposition NP/VP for Synset 1 (See Figure 1) Based on RolesetID

5 Conclusions/Future Directions

This study looked at using the PAS as a pre-processing function to structure unstructured text for evaluation of semantic similarity of two unique sentences. Findings from content analysis showed a way to manage incomplete sentences, justify comparison of differing arguments, and predict pattern for mapping appropriate differing arguments. What remains to be tested is if the methods discovered in this study are repeatable with equivalent synsets from another radiology corpus. It is hypothesized, however, that new predicates would not drastically change the prediction patterns as the distribution or curve (See Figure 3) of predicates is not expected to vary. While a radiology corpus is not entirely representative of the content clinicians will have to review to assure continuum of care, a radiology corpus does provide a good context for NLP experimentation because of its content predictability. Future experiments will have to broaden the clinical corpus content, and those findings based on annotation methods in this study may have different results [26]. Future experiments with this dataset will have to take the annotated PAS corpus and conduct experiments of semantic similarity to determine if the logical representation of the proposition sentence is truly semantically representative of sysnet to accomplish data reduction.

References

1. Zweigenbaum, P., Demner-Fushman, D.: Advanced literature-mining tools. In: Edwards, D., Stajich, J., Hansen, D. (eds.) Bioinformatics, pp. 347–380. Springer, New York (2009)
2. Demner-Fushman, D., Chapman, W.W., McDonald, C.J.: What can natural language processing do for clinical decision support? Journal of Biomedical Informatics 42, 760–772 (2009)
3. Friedman, C., Hripcsak, G.: Natural language processing and its future in medicine. Academic Medicine 74(8), 890–895 (1999)

4. Evashwick, C.: Creating the continuum of care. Health Matrix 7(1), 30–39 (1989)
5. Shortell, S.M., Gillies, R.R., Anderson, D.A.: The new world of managed care: Creating organized delivery systems. Health Affairs 13(5), 46–64 (1994), doi:10.1377/hlthaff.13.5.46
6. Haggerty, J.L., Reid, R.J., Freeman, G.K., Starfield, B.H., Adair, C.E., McKendry, R.: Continuity of care: a multidisciplinary review. BMJ 327(7425), 1219–1221 (2003), doi:10.1136/bmj.327.7425.1219
7. Uijen, A.A., Schers, H.J., Schellevis, F.G., van den Bosch, W.J.H.M.: How unique is continuity of care? A review of continuity and related concepts. Family Practice 29(3), 264–271 (2012), doi:10.1093/fampra
8. Tan, H., Kaliyaperumal, R., Benis, N.: Ontology-Driven Construction of Domain Corpus with Frame Semantics Annotations. In: Gelbukh, A. (ed.) CICLing 2012, Part I. LNCS, vol. 7181, pp. 54–65. Springer, Heidelberg (2012), doi:10.1007/978-3-642-28604-9_5
9. Chou, W.-C., Tsai, R.T.-H., Su, Y.-S., Ku, W., Sung, T.-Y., Hsu, W.-L.: A Semi-Automatic Method for Annotating a Biomedical Proposition Bank, Sydney, Australia. Paper Presented at the Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora (2006)
10. Tsai, R., Chou, W.-C., Su, Y.-S., Lin, Y.-C., Sung, C.-L., Dai, H.-J., et al.: BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. BMC Bioinformatics 8(1), 325 (2007)
11. Cohen, K.B., Hunter, L.: A critical review of PASBio's argument structures for biomedical verbs. BMC Bioinformatics, 7(suppl. 3), S5 (2006)
12. Godbert, E., Royaute, J.: PredXtract, A Generic Platform to Extract in Texts Predicate Argument Structures (PAS), Valleta, Malta. Paper Presented at the LREC 2010 Proceedings (2010)
13. Kilicoglu, H., Fiszman, M., Rosemblat, G., Marimpieti, S., Rindflesch, T.: Arguments of Nominals in Semantic Intepretation of Biomedical Text, Uppsala, Sweden. Paper Presented at the BioNLP 2010 (2010)
14. Kogan, Y., Collier, N., Pakhomov, S., Krauthammer, M.: Towards Semantic Role Labeling & IE in the Medical Literature. Paper Presented at the Annual AMIA Symposium (2005)
15. Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T., Tsujii, J.I.: Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. Paper Presented at the Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia (2006)
16. Surdeanu, M., Harabagiu, S., Williams, J., Aarseth, P.: Using Predicate-Argument Structures for Information Extraction. Paper Presented at the Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan (July 2003)
17. Tsai, R., Chou, W.-C., Su, Y.-S., Lin, Y.-C., Sung, C.-L., Dai, H.-J., Hsu, W.-L.: BIOSMILE: A semantic role labeling system for biomedical verbs using a maximum-entropy model with automatically generated template features. BMC Bioinformatics 8(1), 325 (2007)
18. Wattarujeekrit, T., Shah, P., Collier, N.: PASBio: Predicate-argument structures for event extraction in molecular biology. BMC Bioinformatics 5(1), 155 (2004)
19. Samsonovic, A.V., Ascoli, G.A.: Principal semantic components of language and the measurement of meaning. PLoS ONE 5(6), e10921 (2010)
20. Caviedes, J.E., Cimino, J.J.: Towards the development of a conceptual distance metric for the UMLS. Journal of Biomedical Informatics 37(2), 77–85 (2004)

21. Chaves-González, J.M., Martínez-Gil, J.: Evolutionary algorithm based on different semantic similarity functions for synonym recognition in the biomedical domain. Knowledge-Based Systems 37, 62–69 (2013), doi: http://dx.doi.org/10.1016/j.knosys.2012.07.005
22. Builtelaar, P., Sacaleanu, B.: Ranking and Selecting Synsets by Domain Relevance. Paper Presented at the Proceedings of WordNet and Other Lexical Resources (2001)
23. Elhadad, N., Sutaria, K.: Mining a Lexicon of Technical Terms and Lay Equivalents. Paper presented at the Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing, Prague, Czech Republic (2007)
24. Mihalcea, R., Corley, C., Strapparava, C.: Corpus-Based and Knowledge-Based Measures of Text Semantic Similarity. Paper Presented at the Proceedings of the 21st National Conference on Artificial intelligence, Boston, Massachusetts (2006)
25. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications. Journal of the American Medical Informatics Association 17(5), 507–513 (2010), doi:10.1136/jamia.2009.001560
26. Albright, D., Lanfranchi, A., Fredriksen, A., Styler, W.F., Warner, C., Hwang, J.D., Savova, G.K.: Towards comprehensive syntactic and semantic annotations of the clinical narrative. Journal of the American Medical Informatics Association (2013), doi:10.1136/amiajnl-2012-001317
27. Friedlin, J., Mahoui, M., Jones, J., Jamieson, P.: Knowledge Discovery and Data Mining of Free Text Radiology Reports. In: 2011 First IEEE International Conference on Paper Presented at the Healthcare Informatics, Imaging and Systems Biology, HISB, July 26-29 (2011)
28. McCrae, J., Collier, N.: Synonym set extraction from the biomedical literature by lexical pattern discovery. BMC Bioinformatics 9(159) (2008)
29. Xia, F., Yetisgen-Yildiz, M.: Clinical Corpus Annotation: Challenges and Strategies, Istanbul, Turkey. Paper Presented at the Third Workshop on Building and Evaluating Resources for Biomedical Text Mining Workshop Programme (2012)
30. Babko-Malaya, O.: Propbank Annotation Guidelines (2005), http://verbs.colorado.edu/~mpalmer/projects/ace/PBguidelines.pdf (retrieved November 7, 2010)
31. Unified Verb Index (2012), http://verbs.colorado.edu/verb-index/index.php (retrieved December 12, 2012)
32. Yu, C.H., Jannasch-Pennell, A., DiGangi, S.: Compatibility between text mining and qualitative research in the perspectives of grounded theory, content analysis, and reliability. The Qualitative Report 16(3), 730–744 (2011)
33. Holden, R.J.: Physicians' beliefs about using EMR and CPOE: In pursuit of a contextualized understanding of health IT use behavior. International Journal of Medical Informatics 79(2), 71–80 (2010)