## Domain Role Ontologies

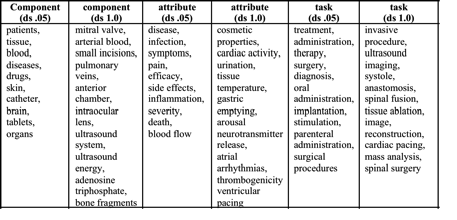
In contrast with the invention keyterms, which focus on roles of terms within individual patents, the purpose of domain role ontologies is to capture broad but useful distinctions within term vocabularies across entire domains. Specifically, this module divides terms into *tasks*, *components*, and *attributes*. Tasks represent activities performed or goals achieved; components capture the objects comprising, produced or acted upon; and attributes refer to the dimensions or metrics that inventions seek to affect. Attributes are further classified with respect to their default *polarity* (whether the dimension is one typically intended to be increased or decreased).

For example, within the computer science domain, tasks would include terms like *backup operation, document management, communication*; components would include *memory subsystem, source file, object access protocol*; attributes include *cpu utilization, transport layer security, performance bottleneck*.

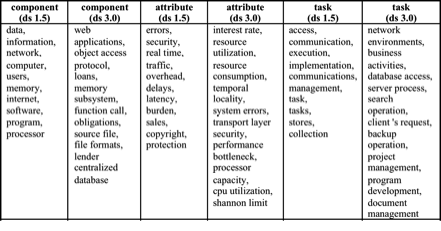
The patent corpus is divided into subcorpora by domain and by time (year of application). For each subcorpus, a domain-specific naïve Bayes classifier is automatically generated using bootstrapping from a general-purpose annotation set. The domain specific classifier is then used to categorize terms for that domain. Additionally, a *domain specificity score* is computed for each term, based on the probability of appearance within a randomized subset of the entire patent collection. This allows an analyst to browse the categorized terminology for any domain at varying levels of domain specificity.

Models have been constructed for the domains of computer science, health, chemistry, agriculture and construction. Examples of output for the health domain and computer science domain (at two levels of domain specificity) are presented below.

Health classification examples;



Computer science examples:



Examples of polarity classification for the health and computer science are shown below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Domain** | **# attributes** | **% of total** | **Examples** |
| *health (pos)* | 43807 | 54% | ambulation, hemodynamic performance, atrial rate, anticoagulant activity, coaptation, blood oxygen saturation |
| *health (neg)* | 37063 | 46% | bronchospasm, thrombogenicity, ventricular pacing, withdrawal symptoms, fibrin formation, cardiac dysfunction |
| *computer science (pos)* | 32291 | 44% | transport layer security, processor capacity, cpu utilization, routability, network speeds, microprocessor performance |
| *computer science (neg)* | 40998 | 56% | identity theft, deadlocks, system overhead, memory fragmentation, risk exposure, bus contention, software development costs, network latencies, data entry errors |