**Annotation Guideline**

## Description of Annotation Task

This annotation guideline is about labeling issues in task of citation recommendation based on argumentative zoning of user queries. To make full use of semantic information containing in user query, we need to construct a dataset with pairs of labeled citing sentences and corresponding citation. This guideline mainly introduces methods that how to label each user query with a specific argumentative category defined in this paper.

When investigating categories of user query, we conduct the literature survey about scientific text argumentative zoning and citation function classification. Our scheme is derived from these two areas of research. By further pre-labeling the sample data, we proposed our classification scheme shown in Table 1 below:

Table 1 Classification Scheme of User Queries

|  |  |
| --- | --- |
| Category | Description |
| Method | A description of the methods used in the research process. The methods here can refer to information such as experimental methods, procedures, data, information resources, tools, parameters, formats, standards, protocols, models, etc. |
| Conclusion | Conclusions based on experimental results and experimental phenomena. |
| Goal | A description of the research goal or an introduction to the research. |
| Object | Explanation or introduction of a research object or topic. |

## Rules for Argumentative Category Recognition

The formulation of annotation rules is an important step before data annotation. It can not only give the definition of each category, but also provide guidance and assistant for annotators to standardize and improve the quality and efficiency of annotation process. Before labeling, we firstly give the specific labeling rules for each category.

Basically, there are two kinds of cues for annotators to use. First is the semantic information of citation context. For each citing sentence, we replace those citation placeholders with string *[CITE]*. Therefore, annotators can make full usage of contextual information during labeling. The second cue is clue words and sentence templates. For each argumentative category, we tried to list some typical words and sentences pattens which can be clues for annotation. The given suggestion may be incomplete, annotators need to make judgements based on their own situation. When labeling categories of Conclusion, Goal and Object, and distinguishing between two different categories that containing in the same sentence, we suggest annotating the data mainly based on the sentence templates.

Normally, there should only be one category for one citing sentence. If the citing sentence covers several categories, which can not be classified clearly based on the annotation guideline, we ask the annotator to skip the sentence.

## Description and Examples for Different Categories

This part will give description, rules and examples for each argumentative category proposed in this paper. Annotators could make judgements based on the clue words and sentences templates.

Clue words here means the words which can bring cues for classification. For example, if the citing sentence contains noun words *method*, word *performed*, and phrase *Spearman's rank correlation analyses*, then annotator would prefer to label it with the *Method* category according to those clue words. Sentence templates refers to the sentence structure which is the way for writers to say what they want.

### Method

If the sentence mentions experimental methods, procedures, data, information resources, tools, parameters, formats, standards, protocols, models and other information, then it can be marked as *Method* category. Clue words are given in Table 2. More detailed examples are given in Table 3.

Table 2 Clue Words for Method category

|  |  |
| --- | --- |
| Word Features | Examples |
| Entity of Known Methods and Tools | Gene knockdown, SAGE：Serial Analysis of Gene Expression, miRGen, miRNA, siRNA, MUSCLE, BLAST, SMART, GeNorm, Rfam, Ensembl, GO, BioCarta, Interpro |
| Entity of Known Database | ModBase, UniProt, GOBASE, KEGG, miRbase, BioGRID |
| Noun and Verb | technique, approach, pathway, protocol, use, apply, utilize |

Table 3 Different kinds of examples for Method category

|  |  |
| --- | --- |
| Descriptions | Examples |
| Technical methods used in the research | 1. As a method verification Spearman's rank correlation analyses of XER and XERcomp against determined blood levels of estradiol and testosterone (total and free) [CITE] were performed on the combined study group data. 2. Rather than monitoring the level of identified xenoestrogens, the integrated biological activity was used to estimate the xenobiotic burden of serum samples using [CITE]. 3. Some missed sequences could be retrieved by adding protein structural information (such as hydrophobic cluster analysis, [CITE]), but these methods require visual expertise and cannot easily be automated. |
| The experiment procedure refers to the operation step information in the experiment | 1. Venous blood samples were collected into 10 ml vacuum tubes and after centrifugation the serum was transferred to Nunc tubes and stored at -80℃ for later analysis [CITE]. 2. Tumor sections were incubated with trypsin at 37℃ for 30 min after deparaffinization, rehydration, and washing as we described recently [CITE]. 3. The level of concordance was measured by periodically re-hybridizing the melanoma cell line A375-melanoma (American Type Culture Collection, Rockville MD) to the reference samples consisting of pooled PBMC as previously described [CITE]. |
| Information resources refer to some databases, standard protocols and other information | 1. A first massive comparison project, OrthoMCL, was designed to cluster malaria genes based on their sequence similarity with genes of 55 other genomes (600,000 sequences), using the BlastP algorithm and Karlin-Altschul E-value statistics to build the all-against-all comparison table [CITE]. 2. Description of gene function follows the guidelines of the Gene Ontology (GO) structured vocabulary [CITE]. 3. The Kyoto Encyclopedia of Genes and Genomes (KEGG) resource provides a set reference of metabolic schemes, manually designed so as to represent all possible primary metabolic reactions and formatted with the KEGG Markup Language (KGML) [CITE]. |
| The software, platforms, systems, and instruments used in the experiment. | 1. The search for ESS with two different programs (Acescan2 and the EBI ASD tools) [CITE] indicated the presence of a high number of ESS in exon 14 in human and mouse FAK gene (Fig and data not shown). 2. The page also supports MUMMER ([CITE]) comparisons between genomes in the database or with a user-supplied sequence. 3. RefSeq ([CITE]) records are used when available to take advantage of their more thorough and consistent annotation. |
| Defined parameters during the experiment, the calculation formula of the index, and the description of the calculation process. | 1. A second threshold given by Clb5,6 Cdk1/Sic1 is required, together with the first one, to adjust the Ps value to carbon sources availability (low Ps in poor/ethanol media, high Ps in rich/glucose media) [CITE]. 2. The KetoCal® diet was fed to the mice in paste form (water: KetoCal® 1:2) within the cage using procedures as we previously described [CITE]. 3. For both agents, we used doses below the threshold for growth retardation and systemic toxicity ([CITE]): 1 mg/kg for chlorpyrifos and either 1 or 2 mg/kg for diazinon. |
| Subjects used in the experiment | 1. A DL4-IgG fusion protein was used to monitor Notch1 expression ([CITE]). 2. siRNA duplexes (SMARTpool, Dharmacon Research) designed to repress GSK-3; or non-specific control pool duplexes or siRNA duplexes designed to target lamin B1 were transfected into DLD-1 cells using OligofectAMINE; (Invitrogen) as describe [CITE]. 3. Preparation for immunochemistry was as described previously [CITE] using rabbit anti-myc (A-14) and anti-GFP (Invitrogen) antibodies, respectively, followed by 10 and 15 nm protein A–gold, respectively (Cell Microscopy Center, Utrecht University, Netherlands). |
| Model name or mechanism with clear information | 1. Since it is well established that the residual familial risk of breast cancer, not caused by BRCA1 or BRCA2 genes, could be explained by a polygenic or high-risk genes heterogeneity model [CITE], we selected individuals affected with breast cancer without mutations in BRCA1/2 genes from high-risk families (one individual per family), in order to increase the power of the study to find genetic variants involved in breast cancer susceptibility. 2. By using animal models, a more optimized study design may be provided [CITE], as a number of identical (inbred) individuals under the same treatment and environmental conditions are compared and tissue collected at the same time point by the same researcher. 3. It is inconsistent with search-and-capture models and computational models based on asters and motors [CITE] because minus ends are not located in discrete nucleating structures at poles. |

### Conclusion

When the sentences contain the conclusions obtained by the cited paper or make summary based on the experimental results in cited paper, annotators could label the citing sentence to be *Conclusion* category. We have clarified three situations: user query contains the description of results and conclusion from cited paper; user query contains the hypotheses and future prospects; user query reveals correlations between two objects. Following clue words in Table 4 can be referred to. More detailed examples are given in Table 5.

Table 4 Clue Words for Method category

|  |  |
| --- | --- |
| Word Features | Examples |
| Verb | report, verify, identify, cause, induce, found, seen, suggest, reveal, indicate, demonstrate, support |
| Nouns, adjectives, prepositions, adverbs | evidence, associated, correlated, important, related, empirically, future, hypothesis, importance, association, thus, because, since |
| Phrases | indeed, in fact, result in, lead to, consistent with, be regarded as, be believed to |

Table 5 Different kinds of examples for Conclusion category

|  |  |
| --- | --- |
| Descriptions | Examples |
| Experimental results and experimental phenomena | Apparent HIV-1 assembly in endosomes appears especially prominent in macrophages, in which very large numbers of virion particles are found within compartments containing the late endosomal markers CD63 and lysosome-associated membrane protein-1 (Lamp1) [CITE]. |
| The description of experimental conclusions of cited articles which often contains clue words. | 1. In support to our data, it was recently reported that high levels of PCBs in Slovakia male serum samples were associated with a decreased ER mediated activity and increased AhR mediated activity [CITE]. 2. Since the failure of treatment with corticosteroids to ameliorate severe cerebral malaria has been used as evidence against cytokine involvement, it warrants recalling that MIF, known to be high in this circumstance [CITE], antagonizes glucocorticoids, and nitric oxide (noting iNOS is also high [CITE]) inhibits glucocorticoid binding to its receptor. |
| The hypothesis and the outlook for the future. | 1. If it is indeed because of a requirement for a more stable secondary structure, then we may expect that the predicted stability of mRNAs from closed regions would be greater than those in open [CITE]. 2. Future studies are required to directly assess the link between NeuroD6 expression and the rate of fusion and fission using a mitochondrial matrix-targeted photoactivable green fluorescent protein [CITE]. |
| Discussion about whether there is a relationship between two research objects. The following relationships can be mentioned: positive correlation, negative correlation, related, irrelevant, cause, help, etc. | 1. A high degree of correlation exists between the in vitro and in vivo transcriptomes of P.falciparum with an overexpression seen for genes encoding a sexual stage antigen as well as gene families that encode surface proteins, providing interesting new vaccine candidates [CITE]. 2. In addition, TII is not associated with patient' gender, tumor location, or growth pattern [CITE]. 3. In murine neonatal thymus, TSLP produced from medullary thymic epithelia cells (mTEC) contributes to the expression of FoxP3 and the maturation of natural regulatory T cells [CITE]. |

### Goal

When make judgement about this category, we suggest referring to two sentence templates, one is “Someone did something”, another one is “Someone did something for a certain purpose”. This category is derived from classification schema of Argumentative Zoning and CoreSC. We list the similar category definitions described in different papers in Table 6.

Table 6 Relevant Definitions for Goal category

|  |  |  |
| --- | --- | --- |
| Related Study | Similar Category | Category Definition |
| Argumentative Zoning I | Aim | Specific research goal of the current paper |
| Argumentative Zoning II | Aim | Statement of specific research goal, or hypothesis of current paper |
| CoreSC | Motivation | A target state of the investigation where intended discoveries are made |

Based on the guideline, citation examples of *Goal* category are given below.

1. For the present study the aim was to recruit 300 male spouses to pregnant women in Greenland, Warsaw (Poland) and Kharkiv (Ukraine), and a subgroup (n = 100) of an already established cohort of fishermen from Sweden was also included in the study [CITE].
2. For instance, Masuda and colleagues [CITE] compared the molecular profile of the same RA FLSs cultured at low density (proliferating) and high density (quiescent).
3. We have previously compared the systemic effects of rIL-2 in peripheral blood monocytes (PBMC) to the peripheral effects induced in melanoma metastases three hours after administration to patient with metastatic melanoma [CITE].

### Object

When the sentence is the description or introduction for a certain research object or topic, the annotator can mark the citation sentence as an *Object* category. This category is based on the CoreSC schema. Relevant categories in CoreSC are given in Table 7.

Table 7 Relevant Definitions for Goal category in CoreSC schema

|  |  |
| --- | --- |
| Similar Category | Category Definition |
| Object-New | An entity which is a product or main theme of the investigation |
| Object-New-Advantage | Advantage of an object |
| Object-New-Disadvantage | Disadvantage of an object |

Normally, when the author wants to describe the definition or composition of some research object, there will be some specific clue words, such as *contain* or i*s a member of*. However, the attributes description of research object is relatively broad. The authors can not only explain the importance and necessity of the object, but also introduce the related background knowledge. Following clue words in Table 8 can be referred to. More detailed examples are given in Table 9.

Table 8 Clue Words for Method category

|  |  |
| --- | --- |
| Word Features | Examples |
| Verb | contain |
| Nouns, adjectives, prepositions, adverbs | necessary, required, essential |
| Phrases | is a member of, play a crucial role |

Table 9 Different kinds of examples for Object category

|  |  |
| --- | --- |
| Descriptions | Examples |
| Definition of the research object, sometime author will give more information, such as the object belongs to a certain type or has certain structure, etc. | 1. CISH is a member of the SOCS adaptor family associated with TCR-mediated proliferation and survival of T cells [CITE] and its expression is specifically dependent upon the activation of STAT-5). 2. MCAK is a member of the Kinesin-13 family [CITE], whose members depolymerize microtubules rather than translocate along them. 3. PINCH is an evolutionarily conserved adapter protein and has five LIM domains [CITE]. |
| Explanation of the research object composition. | 1. In general, microtubules in cells turn over much more rapidly than microtubules assembled from pure tubulin in vitro, due to cellular factors that contribute to increased microtubule turnover, including Op18, Tog, and the microtubule depolymerizing kinesins, MCAK and Kif2A [CITE]. 2. This kinase-like domain is able to interact with several components of cell-matrix contact sites including CH-ILKBP (-parvin, actopaxin), 1, 2, and 3 integrin cytoplasm tails, -parvin (affixin), and paxillin [CITE]. |
| The description of the research object attributes. The attributes here include the nature and relationship of the object. Common descriptions include: the research object is located in a certain position, the research object has a certain function, and the research object has certain characteristics. In addition, when explaining the necessity and importance of the research object, it can be judged that the attribute of the research object is being discussed, that is, the sentence is marked as the object class. | 1. Indeed, in a caveolin-1 (cav-1)-deficient cell line (human hepatoma 7) and embryonic fibroblasts from a cav-1 knockout mouse, SV40 exploits an alternative, cav-1-independent pathway and this alternative pathway is also available in wild-type embryonic fibroblasts [CITE]. 2. In a second TGN to ER pathway, the lipid-sorted pathway utilised by STx traffics from the TGN to the ER in a COP-I independent manner, in a manner controlled by Rab6 [CITE]. 3. One popular cell type is the marsupial PtK2 cell, from the kidney of a normal adult male Potorous tridactylis, which has a large flat morphology and a small number of large chromosomes (for example see [CITE]). 4. Both Abl family kinases and cortactin localize to epithelial adherens junctions where they may similarly interact to promote actin assembly critical to their stability [CITE]. |
| Sentences mention the research object and the citation mark closely follows the research object. | 1. The second complex includes the PDZ protein PAR-3, which binds to atypical protein kinase C (aPKC) [CITE], and the PDZ protein PAR-6. 2. A suggested pH-dependent scaffolding by NHE1 is also a putative mechanism because NHE1 binds PI(4,5)P2 [CITE] and the ezrin-radixin-moesin protein ezrin, and ezrin is suggested to sequester Dbl to plasma membrane microdomains and to regulate Dbl activation of Cdc42. |

## Considerations of Category Priority

Since citation sentences often involve multiple categories of information, this section explains the category priorities that may exist in the labeling process. We recommend that the annotators follow a certain order when labeling. Firstly, they need to label the whole dataset based on only one category. Then, they continue to label the remaining data sequentially according to the left categories. The suggested category order is method, conclusion, goal and object.

### Method and Conclusion

When a clear method and conclusion appear in the citation sentence at the same time, the citation sentence is marked as a *Method* category. Example are given below:

1. Even in multivariate analyses, MVD is related to survival in the whole group of patients with CRC [CITE] or subgroups of patients with stage II-III or stage A-C tumours.
2. Apparently not, as 53BP1 plays no role in this process when tested with in an I-Sce I-based system developed in the Jasin laboratory that measures gene conversion [CITE].
3. In searches of the C. reinhardtii proteome database, CaM-IP2 is flagellar-associated protein 91 (FAP91; [CITE]).

### Method and Object

When the object described in the citation sentence belongs to the *Method* category, the citation sentence is marked as a *Method* category. Example are given below:

1. PlasmoDB operates inside ApiDB, a master web portal for apicomplexan genomes [CITE].
2. PubChem, a repository for molecules acting on biological targets was recently launched, and the UniProt protein knowledge base was recently upgraded to report toxic doses of small molecules on proteins [CITE], however these initiatives are just starting points.
3. Because RASF are known to be one of the key mediators of cartilage destruction in RA, comprehensive data have emerged in recent years from gene expression analyses identifying diagnostically and therapeutically highly valued pathophysiological targets of RASF that mediate joint destruction and inflammation [CITE].

### Method and Goal

When the method and the goal appear in the citation sentence at the same time, the citation sentence is marked as the method category. The sentence patterns that can be referred to are "use a certain method to do " and "use a certain method to do something for a purpose". Example are given below:

1. Zimmermann and colleagues [CITE] used negative selection with anti-CD14 magnetic beads to obtain a relatively clean population of RA FLSs (passage 1).
2. Recently, Koralov et al generated genetically modified mice to study VH replacement [CITE].
3. Nguyen and colleagues studied the relation between peptide YY and cholecystokinin concentrations and gastric emptying in 39 mechanically ventilated intensive care unit patients, two-thirds of whom presented with delayed gastric emptying as assessed with the 13C-octanoate breath test [CITE].

### Conclusion and Method

The method category covers multiple subcategories. When some citation sentences involve the "data" subcategory in the method category in the process of describing the conclusion, the citation sentences are marked as the conclusion category. Example are given below:

1. Those data are consistent with previous data in the mouse ([CITE]) but not with others reported for human DCs, indicating an inability of the hyphal form to induce IL-12.
2. In a population-based study of Swedish women 50-59 years of age (n = 820), low levels of blood lead (mean 2.2 μg/dL; 5th-95th percentiles, 1.1-4.6 μg/dL) were inversely correlated with creatinine clearance and glomerular filtration rate, after adjusting for age, body mass index, urinary or blood cadmium, hypertension, diabetes, and regular use of nonsteroidal anti-inflammatory drug (NSAID) medication ([CITE]).
3. Recent genetic data support this view of sequential activation [CITE].

### Conclusion and Goal

When the citation sentence contains the clue words given in the conclusion category, but the citation sentence actually belongs to the Goal category, the citation sentence is marked with the Goal category. Example are given below:

1. The genes encoding all its subunits have been identified in Drosophila and some of them have been functionally assessed [CITE].

### Object and Method

When a citation sentence describes an object and conforms to the object sentence pattern template, although the citation sentence mentions method-type clue words, the citation sentence is marked as the object category. Example are given below:

1. By using a synthetic peptide encoding the C-terminal 52-96 amino acid sequence of Vpr, which contains the apoptosis inducing domain [CITE], we investigated the effect of Vpr in primary human monocytes and in THP1 promonocytic cell line.

### Goal and Method

When the citation sentence conforms to the goal sentence pattern template, although the citation sentence contains method clue words, the citation sentence is marked as the goal category. Example are given below:

1. In the present work, over 20,000 clones were sequenced from a normalized cDNA library from G.morsitans morsitans and analyzed in combination with previously described ESTs from adult tsetse fat body and midgut [CITE], permitting a uniquely detailed analysis of the sialome of a haematophagous Dipteran.