# P, R, F值总计

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Ftag** | **ManCE qualified** | **SysCE** | **SysCE rigidly matched** | **P\_e** | **R\_e** | **F\_e** |
| f0001 | 71 | 144 | 61(entirely) + 2(partly) | 42.36 | 85.92 | 56.74 |
| f0002 | 25 | 42 | 19(entirely) + 2(partly) | 45.24 | 76 | 56.72 |
| f0003 | 129 | 248 | 98(entirely) + 9(partly) | 39.52 | 75.97 | 51.99 |
| f0014 | 30 | 40 | 17(entirely) + 5(partly) | 42.5 | 56.67 | 48.57 |
| f0015 | 25 | 50 | 20(entirely) + 3(partly) | 40 | 80 | 53.33 |
| f0016 | 30 | 48 | 18(entirely) + 3(partly) | 37.5 | 60 | 46.15 |
| f0027 | 17 | 18 | 9(entirely) + 3(partly) | 50 | 52.94 | 51.43 |
| f0028 | 37 | 50 | 24(entirely) + 2(partly) | 48 | 64.86 | 55.17 |
| f0029 | 31 | 31 | 24(entirely) | 77.42 | 77.42 | 77.42 |
| Average |  |  |  | 46.94889 | 69.97556 | 55.28 |

Entirely: 指完全匹配，即所用pattern没有问题，原因结果片段也几乎完整

Partly: 指部分匹配

这里计算的P,R,F 使用entirely匹配的例子来计算。

## 模板完备性 ---- 未召回手工标记例子分析与分类

### 分类说明

* Pattern未写: 这里指某些maintoken未被写成pattern，为了

1. 歧义太多，匹配出很多伪正例，影响模型的正确率

2. 出现概率较低，为避免pattern库过于复杂

* Partly: 指部分匹配，即

1. 原因片段或结果片段不完整

2. 所用pattern匹配有问题：

(1) 或匹配位置不对，如句中有多个maintoken；---- 标记为**p\_1**

(2) 或匹配的pattern不对，同一句可以同时匹配多个pattern，但由于某个pattern P的排序于正确pattern之前，从而匹配时使用了P。---- **p\_2**

(3) pattern写的不够全面，导致某个句子因有些许出入而未匹配。---- **p\_3**

(4) 其他工具的问题，如使用Stanford-parser解析结果不对 ---- **p\_4**

* 这里ManCE qualified, 指人工标记的例子中符合条件的例子：

1. 只在相邻两句内的CELinks ---- 不符合记为**q\_1**

2. 有maintoken的CElinks ---- 不符合记为**q\_2**

### 数据总计

统计9篇文章未召回例子

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 未召回 | pattern没写 | pattern匹配问题 | | | | 不符合pattern匹配要求 | | |
| p\_1 | p\_2 | p\_3 | p\_4 | q\_1 | q\_2 | q1+q2 |
| 477 | 187 | 45 | 5 | 23 | 31 | 5 | 29 | 35 | 14 |
|  |  |  | 2.67% | 12.30% | 16.58% | 2.67% | 15.51% | 18.72% | 7.49% |
|  | 39.20% | 24.064% | 34.225% | | | | 41.711% | | |

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | Pattern未写 | pattern未写比例 |  |  |  | Pattern | 匹配问题 |  |  |  | Pattern匹配比例 |  | 不合格例子 |  |
| Paper ID | 人工标记总数 | 未召回总数 | 未召回比例 |  |  | p\_1 | p\_1比例 | p\_2 | p\_2比例 | p\_3 | p\_3比例 | p\_4 | p\_4比例 |  | q\_1 | q\_1比例 | q\_2 | q\_2比例 | q\_1+q\_2 |
| F0001 | 92 | 31 | 33.69565217 | 8 | 25.80645161 | 1 | 0.032258 | 1 | 0.032258 | 0 | 0 | 0 | 0 | 0.064516129 | 11 | 0.354839 | 5 | 0.16129 | 5 |
| F0002 | 28 | 9 | 32.14285714 | 2 | 22.22222222 | 0 | 0 | 2 | 0.222222 | 1 | 0.111111 | 1 | 0.111111 | 0.444444444 | 1 | 0.111111 | 2 | 0.222222 | 0 |
| F0003 | 145 | 47 | 32.4137931 | 12 | 25.53191489 | 3 | 0.06383 | 11 | 0.234043 | 5 | 0.106383 | 2 | 0.042553 | 0.446808511 | 6 | 0.12766 | 5 | 0.106383 | 3 |
| F0014 | 36 | 19 | 52.77777778 | 3 | 15.78947368 | 1 | 0.052632 | 3 | 0.157895 | 6 | 0.315789 | 1 | 0.052632 | 0.578947368 | 2 | 0.105263 | 2 | 0.105263 | 1 |
| F0015 | 34 | 14 | 41.17647059 | 3 | 21.42857143 | 0 | 0 | 0 | 0 | 2 | 0.142857 | 0 | 0 | 0.142857143 | 3 | 0.214286 | 6 | 0.428571 | 0 |
| F0016 | 35 | 17 | 48.57142857 | 3 | 17.64705882 | 0 | 0 | 2 | 0.117647 | 8 | 0.470588 | 0 | 0 | 0.588235294 | 2 | 0.117647 | 1 | 0.058824 | 1 |
| F0027 | 23 | 14 | 60.86956522 | 3 | 21.42857143 | 0 | 0 | 1 | 0.071429 | 4 | 0.285714 | 0 | 0 | 0.357142857 | 0 | 0 | 4 | 0.285714 | 2 |
| F0028 | 48 | 24 | 50 | 6 | 25 | 0 | 0 | 2 | 0.083333 | 4 | 0.166667 | 1 | 0.041667 | 0.291666667 | 3 | 0.125 | 6 | 0.25 | 2 |
| F0029 | 36 | 12 | 33.33333333 | 5 | 41.66666667 | 0 | 0 | 1 | 0.083333 | 1 | 0.083333 | 0 | 0 | 0.166666667 | 1 | 0.083333 | 4 | 0.333333 | 0 |
| Total | 477 | 187 | 42.7756531 | 45 | 24.0578812 | 5 | 0.016524 | 23 | 0.111351 | 31 | 0.186938 | 5 | 0.027551 | 0.342365009 | 29 | 0.137682 | 35 | 0.216845 | 14 |

## F0001

人工标记92，共61例完全召回

**---------------------------------------------------------------------------------------------------------------**

**Case1:** “with” 没写成pattern

**With rapid and continual expansion of texts, pictures and videos in cyberspace**, automatic summarization becomes more and more desirable.

**Case5:** “with” 没写成pattern

**With the development of sciences, researchers are limited in time and energy to read more and more publications.** Researchers have to focus on the literature within recent years.

**Case75:** “with” 没写成pattern

There have been more and more real requirements of summarizing pictures **with the explosion of digital pictures online in recently years** due to the wide deployment of cameras and popularity of smart phones.

**Case12:** “in” 没写成pattern

Any method is limited in **its inventor’s knowledge and understandings of problems.**

**Case42:** “to” 没写成pattern

**To reflect not only reading characteristics but also understanding characteristics,** it is a reasonable method to combine the linear organization with the order of generalization and specialization.

**Case69:** “to” 没写成pattern

The patterns of events, conversations, and behaviors should be characterized **to get semantically meaningful summaries of complicated video contents [Tewfik, 1999].**

**Case66:** “In this way” 没写成pattern

**Existing summaries can be put into the space and linked to the tag set.** In this way, the existing summaries can be reused when making new summarizations.

**Case83:** “From this point of view” 没写成pattern

**Summarization is a special transformation that operates dimension reduction for easier understanding.** From this point of view, summarization can be regarded as a transformation of reducing the dimensions of a representation so that the dimensions of representation can be linked to and merge with the dimensions in the mental space.

-------------------------------------------------------------------------------------------------------

**Case52:** such as 与 as 同时出现，as的pattern未匹配p\_1

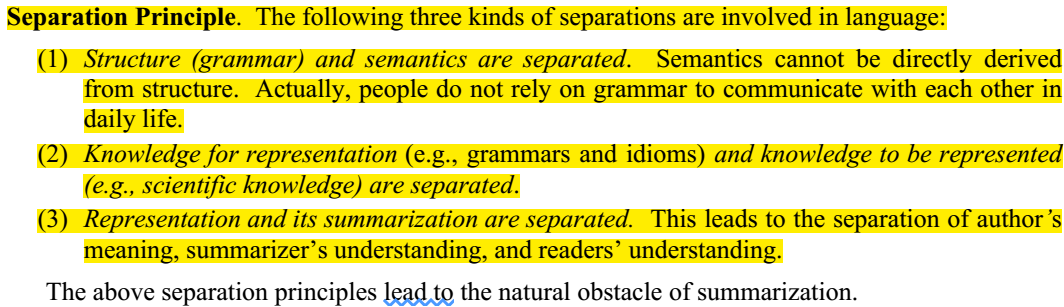
On one hand, a very simple and short representation such as a short paragraph of text and a picture is unnecessary for summarization **as they can be quickly understood.**

**Case71:** provide 与 for 两个pattern，匹配了for的pattern p\_2

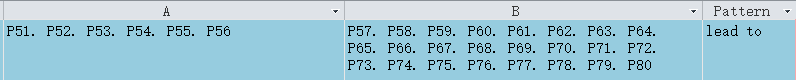
**Current online movies contain subtitles**, which provide a new condition for making summarization of videos through natural language processing.

-------------------------------------------------------------------------------------------------------

**Case16:** 多句 q\_1



**Case17:** 多句 q\_1



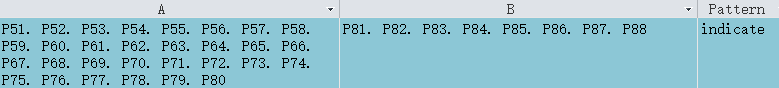
**Case19:** 多句q\_1



**Case23:** 多句 q\_1



**Case26:** 多句 q\_1



**Case27:** 多句q\_1



**Case28:** 多句q\_1



**Case36:** 多句 q\_1

**Humans have been composing complex representations and making summarization through times, so we have the following axiom.** Axiom (Additive Axiom). *A representation can be composed by a set of representations*.

**Case37:** 句子不相邻 q\_1

**Axiom (Additive Axiom). *A representation can be composed by a set of representations*.** This axiom is the basis of representation (including using languages) and summarization (especially, for multi-document summarization). Therefore, a representation *p* can be formalized as a structure of representations: *p*=*p* (*p*) | *p* ∪… ∪*p* | {*p*, …, *p*}, which represents a recursive structure of an abstraction *p* (*p*), an union *p* ∪… ∪*p*, and a set of representations {*p*, …, *p*}

**Case64:** 多句q\_1

**The core words such as “CIKM2012”, “hotel”, “golf” and “garden” can be identified by comparing the source text and tags. Then, the relations like “back of” relying on the core words can be identified.** So, the techniques of text summary can be extended to the construction of semantic link network and image retrieval [Gudivada and Raghavan, 1995].

**Case68:** 多句q\_1

It is important in video management, retrieval and browsing. It becomes more and more  
important with wide use of digital cameras in our society for security, news, entertainment, education, advertisements, etc. **Humans are able to make operations to summarize a movie according to their understandings and requirements, but are limited in ability to view and summarize huge volumes of videos generated everyday. Automatic video summarization can help humans quickly know the key content in big video volumes.**

**-----------------------------------------------------------------------------------**

**Case2:** Implicit cause-effect link q\_2

Text summarization has been studied for over half century, but it is still hard to automatically generate a satisfied summary. **Traditional methods process texts empirically and neglect the fundamental characteristics and principles of language use and understanding.**

**Case14:** Implicit cause-effect link q\_2, 这里as是“作为”, 而不是“因为”。

**As a kind of language representation**, summarization should be based on the basic principles of language use.

错误匹配：

Pattern: 26 [['as']] ---- [['&R@Complete@', '(,)', '(-such/-same/-seem/-regard/-regards/-regarded/-view/-views/-viewed/-denote/-denoted/-denotes)'], ['(-if/-follow/-follows/-&adv)', '&C@Complete@']]

sentTXT: This paper attempts to explore the problem of summarization from multiple dimensions -LRB- especially from empiricism , rationalism , evolutionism , and social/individual constructionism -RRB- so as to form a general summarization methodology . As a kind of language representation, summarization should be based on the basic principles of language use.

Cause: a kind of language representation, summarization should be based on the basic principles of language use.

Effect: This paper attempts to explore the problem of summarization from multiple dimensions -LRB- especially from empiricism, rationalism, evolutionism , and social/individual constructionism -RRB- so as to form a general summarization methodology .

**Case31:** Implicit cause-effect link q\_2

**Some implicit citations have mark words such as “someone says” and according to “someone’s opinion”.** These implicit citations can be located and transformed into explicit citation by searching these mark words and the references according to author names mentioned in text and then inserting uniform citation marks like scientific papers.

**Case43:** Implicit cause-effect link q\_2

**Humans experience in a multi-dimensional space but have to use a two-dimensional media such as paper and screen to externalize representation. Information loses through transformation from the internal representation to the form of display.** Inventing a new interface that can easily convey representations through multiple channels is a way to improve human understanding.

**Case86:** Implicit cause-effect link q\_2

The cognitive level for a particular research field can be reflected by all of its papers. **It stands for the basic cognitive level of all the authors in the field.**

**-----------------------------------------------------------------------------------**

**Case32:** Implicit cause-effect link q\_2, 多句q\_1

**Some implicit citations just reuse others’ sentences or clauses without any mark word. For example, the sentence “It’s Greek to me!” appeared in text implicitly cite the scenario used in 1599 in Shakespeare's play Julius Caesar, and “into thin air” cites the book written by Jon Krakauer published in 1997 and many other earlier works.** Transforming implicit citations needs to compare common clauses in works published in different times.

**Case4:** Implicit cause-effect link q\_2, 多句q\_1

**Summarization carries out in an interaction environment where people read, write, cite and communicate with personal spaces. The personal spaces reflect personal reading experience, interests and knowledge based on the texts that have been read.** A summarization can be satisfied only when it matches the personal space of the reader.

**Case46:** Implicit cause-effect link q\_2, 多句 q\_1

New interface devices like 3D monitor will significantly influence the representation of summary. **Optimizing the layout of display needs to consider three dimensions. 3D printer extends display from cyberspace to the physical space.**

**Case57:** Implicit cause-effect link q\_2, 多句 q\_1

The summary with pictures is more attractive than the text-only summary. **A picture can convey meaning in about 1-10 seconds due to its familiarity and complexity to the viewer. In contrast, readers need to scan the whole text to know the text-only summary.**

**Case79:** Implicit cause-effect link q\_2, 多句 q\_1

**Humans have been pursuing the ways to represent thoughts, behaviors, artifacts and the nature. Various devices and approaches have been invented and developed to represent and process different forms such as natural languages, pictures, videos and graphs. Various representations constitute a representation space with particular structure and operations.** Summarization is a kind of operation that inputs one or more representations in this space and then outputs a new representation.

**-----------------------------------------------------------------------------------------------------------------**

统计未召回例子

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 未召回 | pattern没写 | pattern匹配问题 | | | | 不符合pattern匹配要求 | | |
| p\_1 | p\_2 | p\_3 | p\_4 | q\_1 | q\_2 | q1+q2 |
| 92 | 31 | 8 | 1 | 1 | 0 | 0 | 11 | 5 | 5 |

## F0002

人工标记28，共19例完全召回

**---------------------------------------------------------------------------------------------------------------**

**Case5:** “when” 没写成pattern

Knowledge flow spirals are formed **when knowledge flows in a network**.

**Case18:** “Ving by/through Ving” 没写成pattern

This service can help scientists select peers by finding the dense cliques in knowledge flow networks **by looking at the statistical results and distribution of authors, articles, and citations**.

------------------------------------------------------------------------------------------------------

**Case16:** pattern用的不对p\_2

This service can recommend references **by retrieving documents from the Web and digital libraries, ranking them according to their citation rates and the roles of the authors, and tracing and analyzing their citation networks to show the references as a network rather than as a list.**

应该用by的pattern匹配，但是却是according to匹配的，导致原因结果片段不对。

**Case22:** 先被其他pattern匹配了后的片段重新被正确的pattern匹配，导致原因结果片段不全 p\_2

**Based on the evolution of relevant knowledge flow networks**, this service enables scientists to simulate and estimate the development of a discipline, and thus helps them plan research.

该句先被thus匹配后，原因片段“Based on … discipline”才被based on匹配，所以结果片段缺了”and thus helps them plan research”.

**------------------------------------------------------------**

**Case10:** 插入语pattern无法处理 p\_3

**Enhancing inflow, that is, selecting the appropriate higher-energy nodes to cooperate,** is a strategy to increase the efficiency of knowledge flow networking.

**------------------------------------------------------------**

**Case7:** Stanford-parser解析的不对 p\_4

Reflecting the development of a discipline, the distribution of knowledge energy in a network changes **as its citation network evolves**.

Evolves 解析成了名词（NNS），而未解析成动词(VBZ)，导致pattern无法匹配

-------------------------------------------------------------------------------------------------------

**Case14:** 多句 q\_1

**By extracting and mining from scientific data, activities, and documents of an area; analyzing the relationships between results to enrich a knowledge flow network; and tracing its evolution,** an e-science environment can provide scientists with the following services:

“;”被分句，这一段被分为三个句子，不能满足 by的pattern

**----------------------------------------------------------------------**

**Case1:** Implicit cause-effect link q\_2

However, knowledge is dynamic—**it goes through human brains for knowing, invention, propagation, fusion, generalization, and problem solving.**

**Case3:** Implicit cause-effect link q\_2

**Citations between scientific articles imply a knowledge flow from the authors of the article being cited to the authors of the articles that cite it.** The knowledge flow network implicit in the citation network consists of knowledge flows between nodes that process knowledge, including reasoning, fusing, generalizing, inventing, and problem solving, by authors and co-authors.

-------------------------------------------------------------------------------------------------------

统计未召回例子

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 未召回 | pattern没写 | pattern匹配问题 | | | | 不符合pattern匹配要求 | | |
| p\_1 | p\_2 | p\_3 | p\_4 | q\_1 | q\_2 | q1+q2 |
| 28 | 9 | 2 | 0 | 2 | 1 | 1 | 1 | 2 | 0 |

## F0003

人工标记145，共98例完全召回

**---------------------------------------------------------------------------------------------------------------**

**Case6:** “to” 没写成pattern

Reading hypertext has no big difference from reading text in paper although **different front sizes and colors of words can be easily set** to attract readers’ attention.

**Case10:** “to” 没写成pattern

Link structure was analyzed and used **to improve web information retrieval** [42][57].

**Case14:** “to” 没写成pattern

SLN has been developed toward a self-organized semantic networking model **to manage decentralized resources.**

**Case15:** “to” 没写成pattern

An object-oriented SLN language was suggested **to implement the SLN [66]**.

**Case12:** “to do by Ving” 没写成pattern

**To support intelligent applications** by assigning semantic indicators and rules to links and enabling relational reasoning, analogical reasoning, inductive reasoning, and complex reasoning.

**Case65:** “to”没写成pattern

Secondly, **to minimize energy,** behaviors need to be localized.

**Case71:** “when”没写成pattern

Time, place, event, and individual will be linked **when seeking the answer to the question about who, what, where, why, when, and how.**

**Case92:** “when”没写成pattern

The semantic image of an SLN emerges in the mental space **when the symbols link to the sense and to existing semantic images through the physiological space and psychological space.**

**Case40:** “mean that” 没写成pattern

Scientists will be able to access research objects and thoughts as well as their formation processes on demand through times. This means that they can not only communicate with peers but also access important thoughts through time.

**Case103:** ”show that” 没写成pattern

**The above discussion shows that** SLN is different from previous links and graph-based models in goal, model, method and research issues.

**Case142:** “with” 没写成pattern

The ideal of Bush has been realized **with the development of the Web and various advanced Web applications.**

**Case143:** “with” 没写成pattern

The ideal of Gray’s personal memex and world memex will be realized in the near future **with the development of the next-generation Web.**

-------------------------------------------------------------------------------------------------------

**Case20:** pattern匹配位置不对p\_1

**If some soft, light and mobile detection devices can be invented in the future,** many new applications can be imagined, for example, people can get some services while thinking.

“,”匹配位置应该在in the future之后，而不是在soft之后

**Case60:** pattern匹配位置不对 p\_1

**For example, the event of volcanic eruptions in the physical space will** lead to cancelation of relevant flights, and lead to cancelation of the socio events like conferences held within the region.

Lead to 出现两次，而第二次被匹配，从而未能抽取出因果关系。

**Case69:** pattern匹配位置不对 p\_1

**The socio space and the mental space cooperatively create, evolve and reflect the cyber space in real time.** This enables the mental space to reflect more of the other spaces and enables the cyber space to support the complex space.

enable 出现两次，而第二次被匹配，从而未能抽取出因果关系。

**--------------------------------------------------------**

**Case9:** pattern用的不对p\_2

The similar relation can be detected **by analyzing the contents of nodes based on such approaches as the Latent Semantic Analysis (LSA) and the Vector Space Model (VSM)** [46] [62].

应该用by的pattern却优先被based on的pattern先匹配了，导致原因结果片段不对

**Case29:** 句子优先被别的pattern匹配了p\_2

**Fig.1 (c) depicts an important extension of (b): user behaviors can be sensed and fed back to the cyber space for analyzing the patterns of behaviors, and humans can remotely control the actuators to behave in the physical space through the cyber space.** This enables the cyber space to adapt services according to the feedback since behavior change may indicate some psychological change.

本应匹配的This enables的pattern, 但后一句优先被since匹配了, 导致前一片段This enables the cyber space to adapt services according to the feedback无法与前一句凑成一个因果关系。

**Case53:** pattern用的不对p\_2

**If the organs are abstracted as service nodes and the relations between organs are abstracted as semantic links,** an individual physiological space can be abstracted as a complex semantic link network evolving with physiological rules and the flows of material, energy and information through the links.

应该用if的pattern却优先被as的pattern先匹配了，导致原因结果片段不对

**Case54:** pattern用的不对p\_2

**If artifacts are abstracted as semantic nodes with function descriptions, and the relations between artifacts are abstracted as semantic links,** the artifact space can be regarded as a semantic link network evolving with scientific, technological and socio rules.

应该用if的pattern却优先被as的pattern先匹配了，导致原因结果片段不对

**Case55:** pattern用的不对p\_2

**If socio individuals are abstracted as active semantic nodes and socio relations between individuals are abstracted as semantic links,** the socio space can be regarded as a complex semantic link network evolving with socio rules and interactions.

应该用if的pattern却优先被as的pattern先匹配了，导致原因结果片段不对

**Case57:** pattern用的不对p\_2

**If symbols or symbol units (e.g., words, phrases, and sentences) are abstracted as semantic nodes, and the relations (e.g., sequential relation and cause-effect relation) between symbols are abstracted as semantic links**, the symbol space can be regarded as an evolving semantic link network with rules of languages.

应该用if的pattern却优先被as的pattern先匹配了，导致原因结果片段不对

**Case58:** pattern用的不对p\_2

**If information clusters like web pages or function clusters like services are abstracted as semantic nodes and the relation between clusters as semantic links,** the cyber space can be abstracted as a complex semantic link network evolving various interactions.

应该用if的pattern却优先被as的pattern先匹配了，导致原因结果片段不对

**Case68:** pattern用的不对p\_2

As a kind of mirror, the cyber space reflects the other spaces **by recording and linking various interactions in other spaces [89].**

应该用by的pattern却优先被as的pattern先匹配了，导致原因结果片段不对

**Case74:** pattern 用的不对p\_2

Different from previous notions about the mental space [23][47][51], the mental space model depicted in Fig.7 is based on the multi-dimensional classification space, link and the Interactive Semantic Base ISB [89], **since classification, link and interaction are the basis of the development of the mental space.**

应该用since的pattern却优先被based on的pattern先匹配了，导致原因结果片段不对

**Case114:** 句子优先被别的pattern匹配了p\_2

The node with richer types of links takes the higher priority to emerge than that with less or single type of links. **This is because the node with richer types of links offers higher probability to the new link to derive out more links so that communities have higher probability to be enriched or changed.**

本应匹配的This is because的pattern, 但后一句优先被so that匹配了, 导致前一片段This is because the node with richer …无法与前一句凑成一个因果关系。

**Case120:** pattern用得不对p\_2

**For example, linking the agricultural ecological system to the industrial ecological system** enables some waste of industry to be used as the fertilizer of crops, and enables the agricultural products to be the raw materials of some industries.

应当用enables的pattern，但是被as的pattern先匹配了，导致原因结果片段不对。

**--------------------------------------------------------**

**Case47:** pattern写的太死了 p\_3

**As the consequence of collaboration,** the collaborators’ mental spaces evolve toward more commonalities.

写的pattern是as a consequence, 不能处理as the consequence

**Case134**: Pattern写的太死了p\_3

Only the mental space has the ability to emerge and explain socio semantics **since the mental space has been reflecting various spaces and carrying out reasoning in lifetime.**

为了词义消歧，要求since后跟的内容再wordnet中不能表示时间，但是这里lifetime的含义含有time，所以不能被匹配

**Case139:** Pattern写的太死了 p\_3

Extending the hyperlink network to support some preliminary intelligence such as guided browsing, query implicit relations, question answering and explanation **by introducing a semantic space and reasoning mechanisms such as relational reasoning, analogical reasoning, inductive reasoning, and complex reasoning.**

By的pattern写成了&R@complete@ by &Ving@C@, 即要求by之前的片段必须是有主语谓语的句子，这里Extending the hyperlink network to support some…没谓语而没有被匹配

**Case140:** Pattern写的太死了 p\_3

Realizing semantic lens **by integrating with the multi-dimensional classification space.**

By的pattern写成了&R@complete@ by &Ving@C@, 即要求by之前的片段必须是有主语谓语的句子，这里Extending the hyperlink network to support some…没谓语而没有被匹配

**Case144:** since的pattern写的太死p\_3

Exploring the CP3SME will go beyond Turing’s ideal **since traditional machines and the cyber space are limited in ability to realize the CP3SME and cyber-physical-socio intelligence.**

为了词义消歧，要求since后跟的内容再wordnet中不能表示时间，但是这里lifetime的含义含有time，所以不能被匹配

**--------------------------------------------------------**

**Case86:** Stanford-parser 解析不对p\_4

**Regarding semantic nodes of SLN as agents,** multi-agent (MA) techniques can be adopted to enhance SLN [72].

这里“Regarding …. As agents”应当解析为VP（动名词短语），而却解析成了PP（介词短语），导致pattern未被匹配上

**Case87:** Stanford-parser 解析不对p\_4

**Regarding class or object as node,** the object, class, inheritance, polymorphism, method, roles, and functional modeling in OOMD support the specialization of SLN.

这里“Regarding …. As agents”应当解析为VP（动名词短语），而却解析成了PP（介词短语），导致pattern未被匹配上

-------------------------------------------------------------------------------------------------------

**Case31:** 不相邻且多句 q\_1



**Case39:** 多句 q\_1



**Case109:** 多句 q\_1



**Case121:** 多句 q\_1

**Humans can not only sense from multiple channels but also form close loops through sensing, behaving, emerging semantic images, and reasoning. Behaving through one type of channel (e.g., writing) accompanies sensing through another type of channel (e.g., vision). Practice (e.g., dictation) helps establish the links between the behaviors through different channels and between behaviors and semantic images.** This is why blind people cannot write in the normal natural languages. Interactions among various individuals extend the closed loops.

**Case122:** 多句 q\_1

**Humans emerge semantic images in mind, use languages to indicate semantic image, use sketch to describe semantic image, develop a sketch toward a design, and model a design in the cyber space or in the artifact space as a prototype. A prototype can be refined through reflection and reference in the physical space, artifact space, and socio space as well as through reasoning processes in the mental space. The final design can lead to the artifacts in the artifact space through a manufacture process.** This indicates a new design paradigm that enables designers and users to co-experience during the whole design process.

**Case131:** 多句 q\_1

**Psychological research indicates that brain uses different regions to process the first language and non first languages. The epistemology has been emphasized in the effort to solve the semantics problem [77].** This implies that the current approaches to the semantic web, machine translation, and text understanding through processing symbols in the cyber space is questionable in essence.

**--------------------------------------------------------**

**Case28:** Implicit cause-effect link q\_2

Fig.1 (b) depicts the extension from the cyber space to the physical space through various sensors. **Some significant information in the physical space can be automatically sensed, stored and transmitted through the cyber space.**

**Case35:** Implicit cause-effect link q\_2

**Based on above discussion**, the following question naturally emerges: Can links semantically pass through spaces (e.g., physiological space, psychological space, and mental space) to extend machine intelligence and human intelligence?

**Case59:** Implicit cause-effect link q\_2

**A significant event in the physical space** may generate or influence an event in the socio space.

**Case72:** Implicit cause-effect link q\_2

**Different spaces have different types of distance, for example, friendship between individuals, the difference between interests, and the distance between offices.** Reflecting different aspects of the objects needs to use the distances in different spaces.

**Case133:** Implicit cause-effect link q\_2

**New philosophical issues will arise in CP3SME research as the world people live and the way people interact with each other have been changed greatly with the generation and evolution of various spaces.** The scientific pursuit of the CP3SME brings challenging philosophical issues.

**--------------------------------------------------------**

**Case2:** Implicit cause-effect link q\_2, 多句 q\_1



**Case36**: Implicit cause-effect link q\_2, 多句 q\_1



**Case43:** Implicit cause-effect link q\_2, 多句 q\_1



**---------------------------------------------------------------------------------------------------------------**

统计未召回例子

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 未召回 | pattern没写 | pattern匹配问题 | | | | 不符合pattern匹配要求 | | |
| p\_1 | p\_2 | p\_3 | p\_4 | q\_1 | q\_2 | q1+q2 |
| 145 | 47 | 12 | 3 | 11 | 5 | 2 | 6 | 5 | 3 |

## F0014

人工标记36，共17例完全召回

**---------------------------------------------------------------------------------------------------------------**

**Case4:** “in” 没写成pattern

In fact, it has been shown that the existence of g(m; r) is an undecidable question **in that there does not exist a Turing machine which will eventually come to a stop and print a 1 if g(m; r) does not exist.**

**Case13:** “in order to”没写成pattern

**In order to do better than such a purely enumerative method,** it is necessary to use methods which take into account some of the structure of Turing machines.

**Case29:** “in order to”没写成pattern

**In order to get around the fact that all formal systems which are anywhere near adequate for describing recursive function theory are incomplete,** we avoid restriction to any one of them by introducing the notion of a formal theory (not for the first time, of course).

**---------------------------------------------------------------------------------------------------------------**

**Case33:** pattern匹配位置不对 p\_1

This suggests an Alexandrian solution to a knotty question. Perhaps, a machine should be regarded as thinking **if and only if its behavior can most concisely be described by statements, including those of the form “It now believes Pythagoras’ theorem."**

这里if的位置出现两次，匹配的是第一个if，导致后面的句子开头多出了 “and only if” 后面的&C@Complete@无法匹配

**--------------------------------------------------------**

**Case18:** 句子被其他pattern匹配了p\_2

**It will be made at least plausible that a machine with Q internal states and S symbols should be considered as making about 1 2 log QS elementary steps per step of computation and** hence the number of steps in a computation should be multiplied by this factor to get the length of the computation.

应当被hence匹配，但是句子先被as匹配了，导致剩余的片段只有 “and hence the …” 不能被hence匹配。

**Case20:** 句子被其他pattern匹配了p\_2

Any machine can be modified into such a machine **by adding to it facilities for testing a conclusion and having it spend a small fraction of its time trying the integers in order.**

应当被by的pattern匹配，但是却被for的pattern匹配了，导致原因结果不正确

**Case30:** 句子被其他pattern匹配了p\_2

In order to get around the fact that all formal systems which are anywhere near adequate for describing recursive function theory are incomplete, we avoid restriction to any one of them **by introducing the notion of a formal theory (not for the first time, of course).**

应当用by的pattern匹配，却被for的pattern匹配了，导致原因结果不正确。

**----------------------------------------------------**

**Case5:** pattern写的不完善，从句太长了p\_3

In fact, it has been shown that the existence of g(m; r) is an undecidable question **i**n that there does not exist a Turing machine which will eventually come to a stop and print a 1 **if g(m; r) does not exist**.

If的pattern会将其之前的所有内容都人为是结果，导致结果片段太长

**Case7:** pattern写的不完善p\_3

It will therefore find g(m; r) **if it exists**, but will never know enough to give up if g(m; r) does not exist.

If的pattern要求if之前的片段是complete的，即要有主语谓语，而这里只有谓语，不能被匹配

**Case8:** pattern写的不完善p\_3

It will therefore find g(m; r) if it exists, but will never know enough to give up **if g(m; r) does not exist.**

If的pattern要求if之前的片段是complete的，即要有主语谓语，而这里只有谓语，不能被匹配

**Case26:** pattern写的不完善p\_3

Our procedure consists in trying the numbers fk(m; r) in order (again diagonalizing to avoid computations which don’t end.) **This is based on the plausible idea that, in searching for the solution to a problem, the given data should be taken into account.**

Based on的pattern写的不够完善，不能处理这个句型

**Case27:** pattern写的不完善p\_3

The next complication which suggests itself is to revise the order in which recursive functions are considered. One way is to consider ffk(‘)(m; r) diagonalized on k and ‘. This is based on the idea that the best procedure is more likely to be recursively simple, rather than merely to have a low number in the ordering.

Based on的pattern写的不够完善，不能处理这个句型

**Case34:** pattern写的不完善p\_3

**we are imposing our ideas of the structure of the problem on the machine. The point is to do this** so that the procedures we offer are general (will eventually solve every solvable problem) and also are improvable by the methods built into the machine.

这里so that之前的the point is to do this不常见，so that的pattern无法处理，所以找出的原因片段只是 “The point is to do this”, 没有前一句。

**-----------------------------------------------------**

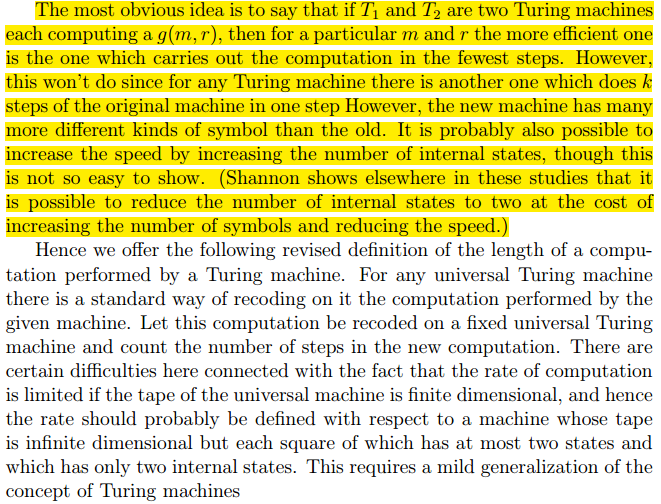
**Case19:** Stanford-parser无法解析特殊符号p\_4

Of course, **if L(m,r,T1) ≤ L(m,r,T2) for all m and r for which these numbers are finite,** we should certainly say that T1 is more efficient than T2.

该句中有特殊符号**≤**, 不能被Stanford-parser解析，无法解析成stanford句法树，无法做pattern匹配

**---------------------------------------------------------------------------------------------------------------**

**Case16:** 多句q\_1

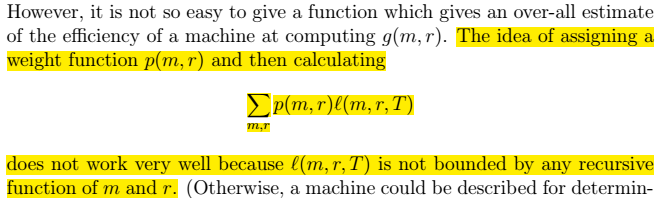


**Case24:** 多句q\_1

**This is known to be a hopeless task. Systems much simpler than Turing machine theory have been shown to have unsolvable decision procedures.** So, we look for a way of evading these difficulties.

**----------------------------------------------------**

**Case21:** Implicit cause-effect link q\_2



**Case25:** Implicit cause-effect link q\_2

This is known to be a hopeless task. **Systems much simpler than Turing machine theory have been shown to have unsolvable decision procedures.**

**----------------------------------------------------**

**Case10:** Implicit cause-effect link q\_2, 多句 q\_1



**---------------------------------------------------------------------------------------------------------------**

统计未召回例子

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 未召回 | pattern没写 | pattern匹配问题 | | | | 不符合pattern匹配要求 | | |
| p\_1 | p\_2 | p\_3 | p\_4 | q\_1 | q\_2 | q1+q2 |
| 36 | 19 | 3 | 1 | 3 | 6 | 1 | 2 | 2 | 1 |

## F0015

人工标记34，共20例完全召回

**---------------------------------------------------------------------------------------------------------------**

**Case12:** “whereby” 没写成pattern

These axioms are used by reasoning programs as part of the process **whereby the program decides what to do.**

**Case29:** “aimed at” 没写成pattern

Daniel Dennett, Tufts University philosopher, has proposed three attitudes **aimed at understanding a system with which one interacts.**

**Case32:** “when” 没写成pattern

It is easiest to understand the ascription of thoughts to machines **in circumstances when we also understand the machine in physical terms.**

**---------------------------------------------------------------------------------------------------------------**

**Case7:** Pattern写的不完善p\_3

The reason for ascribing mental qualities and mental processes to machines is the same as for ascribing them to other people. **It helps understand what they will do, how our actions will affect them, how to compare them with ourselves and how to design them.**

The reason for的pattern无法处理这里的类比句型，导致原因结果片段不对。

**Case18:** Pattern写得不完善p\_3

It is useful **because of its connection with all of these things and because what it says about the dog corresponds in an informative way with similar statements about people.**

Because of的pattern之后接的&NP不作为原因，而这里匹配的&NP一直到句尾，导致原因片段抽取的不对

**---------------------------------------------------------------------------------------------------------------**

**Case8:** 多句 q\_1

Researchers in artificial intelligence (AI) are interested in the use of mental terms to describe machines for two reasons. **First we want to provide machines with theories of knowledge and belief so they can reason about what their users know, don’t know, and want. Second what the user knows about the machine can often best be expressed using mental terms.**

**Case16:** 多句q\_1

C:\Users\Merry\AppData\Local\Temp\1489299181(1).png

**Case28:** 多句q\_1

We suppose that the rules result in a Chinese conversation so intelligent that the person giving and receiving the sentences can’t tell him from an intelligent Chinese. This is analogous to a computer, which only obeys its programming language, but can be programmed such that one can communicate with it in a different programming language, or in English. **Searle says that since the person in the example doesn’t understand Chinese -- even though he can produce intelligent Chinese conversation by following rules -- a computer cannot be said to ‘understand’ things.**

**--------------------------------------------------------**

**Case4:** Implicit cause-effect link q\_2

**However, artificial intelligence is a difficult branch of science and engineering, and, judging by present slow progress**, it might take a long time.

**Case5:** Implicit cause-effect link q\_2

Present machines have almost no emotional qualities, and, in my opinion, it would be a bad idea to give them any. **We have enough trouble figuring out our duties to our fellow humans and to animals without creating a bunch of robots with qualities that would allow anyone to feel sorry for them or would allow them to feel sorry for themselves.**

**Case20:** Implicit cause-effect link q\_2

If the control is placed on a radiator or radiant heated floors, **it will ‘think’ the entire room is hot and will lower your blanket temperature,** making your bed too cold.

**Case26:** Implicit cause-effect link q\_2

**Confronted with an uncomfortable room** you form any of the following hypotheses depending on what other information you had.

**Case30:** Implicit cause-effect link q\_2

The design stance can usefully be applied to a thermostat -- **it shouldn’t be too hard to figure out how to set it, no matter how it works.**

**Case33:** Implicit cause-effect link q\_2

However, we must be careful not to ascribe properties to a machine that  
the particular machine doesn’t have. **We humans can easily fool ourselves  
when there is something we want to believe.**

**---------------------------------------------------------------------------------------------------------------**

统计未召回例子

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 未召回 | pattern没写 | pattern匹配问题 | | | | 不符合pattern匹配要求 | | |
| p\_1 | p\_2 | p\_3 | p\_4 | q\_1 | q\_2 | q1+q2 |
| 34 | 14 | 3 | 0 | 0 | 2 | 0 | 3 | 6 | 0 |

## F0016

人工标记35，共18例完全召回

**---------------------------------------------------------------------------------------------------------------**

**Case5:** “when” 没写成pattern

**When a rule is activated**, MYCIN tests whether the pattern part matches the database.

**Case20:** “not…unless”没写成pattern

An advanced MYCIN might need to know that a patient won’t take a bad tasting medicine **unless he is convinced of its necessity.**

**Case23:** “depend on”没写成pattern

Our ability to use common sense knowledge **depends on being able to do common sense reasoning.**

**---------------------------------------------------------------------------------------------------------------**

**Case3:** pattern用的不对p\_2

Even patients are not really part of the ontology, although MYCIN asks for many facts about the specific patient. **This is because patients aren’t values of variables, and MYCIN never compares the infections of two different patients.**

Because的pattern有好几个，应当用[['because']] ---- [['&R', '(,/./;/--)', '(&AND)', '&THIS', '&BE', '(&ADV)'], ['&C']]，但却被[['because'], [',']] ---- [[], ['&C'], ['&R']]优先匹配

**Case34:** 句子被其他pattern优先匹配了p\_2

We can therefore regard it as a communication convention that if a bird can fly the fact need not be mentioned, but **if the bird can’t fly and it is relevant,** then the fact must be mentioned.

这句话先被therefore的pattern匹配了，结果片段为能被检查

**------------------------------------------------------**

**Case6:** pattern写的不够完善p\_3

**If not the pattern fails** and MYCIN tries another.

if的pattern无法处理这种中间没有 ‘,’ 也没有 ‘then’的句型

**Case12:** pattern写的不够完善p\_3

The most important events are actions, and **for a program to plan intelligently,** it must be able to determine the effects of its own actions.

For的pattern不能处理这种后跟一个名词短语的句型

**Case15:** Pattern写得不完善p\_3

Ignoring prognosis is possible **because of the specific narrow domain in which MYCIN operates.**

Because of的pattern之后接的&NP不作为原因，而这里匹配的&NP一直到句尾，导致原因片段抽取的不对

**Case22:** pattern写的不够完善p\_3

**If MYCIN were extended to be a robot physician** it would have to know common sense physics and maybe also some scientific physics.

if的pattern无法处理这种中间没有 ‘,’ 也没有 ‘then’的句型

**Case26:** pattern写的不够完善p\_3

**If you now learn that the car will be out in half an hour** you reverse yourself again.

if的pattern无法处理这种中间没有 ‘,’ 也没有 ‘then’的句型

**Case27:** pattern写的不够完善p\_3

Several artificial intelligence researchers, for example Marvin Minsky (1974) have pointed out that intelligent computer programs will have to reason non-monotonically. **Some concluded that therefore logic is not an appropriate formalism.**

Therefore的pattern无法识别前面的 some concluded that 片段

**Case28:** pattern写的不够完善p\_3

However, it has turned out that deduction in mathematical logic can be supplemented **by additional modes of non-monotonic reasoning,** which are just as formal as deduction and just as susceptible to mathematical study and computer implementation.

By的pattern没法处理后面不是Ving的句型

**Case33:** pattern写的不够完善p\_3

We can therefore regard it as a communication convention that **if a bird can fly** the fact need not be mentioned, but if the bird can’t fly and it is relevant, then the fact must be mentioned.

if的pattern无法处理这种中间没有 ‘,’ 也没有 ‘then’的句型

**---------------------------------------------------------------------------------------------------------------**

**Case4:** 多句q\_1

**MYCIN’s ontology includes bacteria, symptoms, tests, possible sites of infection, antibiotics and treatments. Doctors, hospitals, illness and death are absent. Even patients are not really part of the ontology, although MYCIN asks for many facts about the specific patient.** **This is because patients aren’t values of variables, and MYCIN never compares the infections of two different patients.** It would therefore be difficult to modify MYCIN to learn from its experience.

**Case35:** 多句q\_1

**Suppose I hire you to build me a bird cage, and you build it without a top, and I refuse to pay on the grounds that my bird might fly away. A judge will side with me. On the other hand suppose you build it with a top, and I refuse to pay full price on the grounds that my bird is a penguin, and the top is a waste. Unless I told you that my bird couldn’t fly, the judge will side with you.** We can therefore regard it as a communication convention that if a bird can fly the fact need not be mentioned, but if the bird can’t fly and it is relevant, then the fact must be mentioned.

**------------------------------------------------------**

**Case8:** Implicit cause-effect link q\_2

The production formalism turned out to be suitable for representing a large amount of information about the diagnosis and treatment of bacterial infections. **When MYCIN is used in its intended manner it scores better than medical students or interns or practicing physicians and on a par with experts in bacterial diseases when the latter are asked to perform in the same way.**

**------------------------------------------------------**

**Case1:** Implicit cause-effect link q\_2 多句q\_1

**Nevertheless, hardly any of them have certain common sense knowledge and ability possessed by any non-feeble-minded human.** This lack makes them “brittle". By this is meant that they are difficult to extend beyond the scope originally contemplated by their designers, and they usually don’t recognize their own limitations.

**---------------------------------------------------------------------------------------------------------------**

统计未召回例子

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 未召回 | pattern没写 | pattern匹配问题 | | | | 不符合pattern匹配要求 | | |
| p\_1 | p\_2 | p\_3 | p\_4 | q\_1 | q\_2 | q1+q2 |
| 35 | 17 | 3 | 0 | 2 | 8 | 0 | 2 | 1 | 1 |

## F0027

人工标记23，共9例完全召回

**---------------------------------------------------------------------------------------------------------------**

**Case20:** “to”没写成pattern

**To encourage data sharing,** it should be rewarded.

**Case4:** “in response to”没写成pattern

**In response to this data deluge,** the systematic use of databases has become an integral part of the scientific process.

**Case17:** “in this way”没写成pattern

**For example, a data-mining paper needs to include the explicit description (database query) of how the data that were analyzed in the paper were collected and filtered, but not the data themselves.** In this way, a reviewer with access to public data could reproduce the data sets and analysis procedures.

**---------------------------------------------------------------------------------------------------------------**

**Case8:** pattern用的不对p\_2

One might argue that complex biological experiments have always been difficult to reproduce, **as there are so many variables**.

应当用as的pattern但优先被so的pattern匹配了。

**--------------------------------------------**

**Case3:** pattern写的太死 p\_3

**As data volumes grow,** it is increasingly arduous to extract knowledge.

As的pattern不能正确匹配这个句型。没有写as &C, &R句型的pattern

**Case7:** pattern写的太死 p\_3

**As experiments yield more data, and analysis becomes more complex,** data become increasingly difficult to document and reproduce.

As的pattern不能正确匹配这个句型。没有写as &C, &R句型的pattern

**Case14:** pattern写的太死 p\_3

**For these data to be useful to others,** they must be published using a controlled vocabulary and in standard forms, and the instruments and measurements must be well specified.

For的pattern只能处理&R for &Ving@C@, 不能处理上述这种句型，歧义太大。

**Case18:** pattern写的太死 p\_3

For the analysis to be repeatable in 20 years’ time **requires archiving both data and tools.**

Requires 的pattern不能处理这个句型，for也不能处理这个句型

**---------------------------------------------------------------------------------------------------------------**

**Case2:** Implicit cause-effect link q\_2

**With data correlated over many dimensions and millions of points, none of the old steps — do experiment, record results, analyse and publish — is straightforward.** Many predict dramatic changes to the way science is done, and suspect that few traditional processes will survive in their current form by 2020 (ref. 1).

**Case6:** Implicit cause-effect link q\_2

Fully documenting these steps would be arduous, and there is little chance that someone could repeat the exact procedure 20 years from now; **both Matlab and GenBank will change enormously in that time.**

**Case10:** Implicit cause-effect link q\_2

A collaboration involving hundreds of Inter net-connected scientists raises questions about standards for data sharing. **Too much effort is wasted on converting from one proprietary data format to another.**

**Case23:** Implicit cause-effect link q\_2

**And so although some sub-disciplines may reach a plateau in data generation, other technological innovations will take their place.** Scientists in 2020 will continue to work in an exponential world.

**-----------------------------------------------------**

**Case9:** Implicit cause-effect link q\_2 多句q\_1

**Increasingly, scientists are analysing complex systems that require data to be combined from several groups and even several disciplines. There are collaborations sharing data across departments and time zones, and important discoveries are made by scientists and teams who combine different skill sets — not just biologists, physicists and chemists, but also computer scientists, statisticians and data-visualization experts.** It is important to realize that today’s graduate students need formal training in areas beyond their central discipline: they need to know some data management, computational concepts and statistical techniques.

**Case22:** Implicit cause-effect link q\_2 多句q\_1

**Exponential growth occurs when a new generation of instruments leapfrogs the previous generation, which become obsolete.** There are two trends in science today, scaling up and scaling out. Some scientists are building billion-dollar facilities, such as astronomy’s Large Synoptic Survey Telescope or the Large Hadron Collider, which are only affordable as international collaborations. Such facilities are not easily leapfrogged. And once these peta-scale experiments are switched on they will produce roughly the same amount of data each year — merely linear growth. But in the scaling-out model, experiments that deploy an array of small instruments can exploit the coming explosion in cheaper commodity technology.

**---------------------------------------------------------------------------------------------------------------**

统计未召回例子

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 未召回 | pattern没写 | pattern匹配问题 | | | | 不符合pattern匹配要求 | | |
| p\_1 | p\_2 | p\_3 | p\_4 | q\_1 | q\_2 | q1+q2 |
| 23 | 14 | 3 | 0 | 1 | 4 | 0 | 0 | 4 | 2 |

## F0028

人工标记48，共24例完全召回

**---------------------------------------------------------------------------------------------------------------**

**Case6:** “to” 没有写成pattern

**To ameliorate these problems,** scientists will need better analysis algorithms that can handle extremely large datasets with approximate algorithms (ones with near-linear execution time) and they will need parallel algorithms that can apply many processors and many disks to the problem to meet cpu density and bandwidth-density demands.

**Case26:** “to” 没写成pattern

**To preserve applications that depend on T,** the database administrator can then define a view over T’ and T” corresponding to the original definition of table T, allowing old programs to continue to operate correctly.

**Case10:** “when”没写成pattern

**When a scientist wants to correlate data from two different data centers**, then there is no option but to move part of the data from one place to another.

**Case14:** “in order to”没写成pattern

Data must be self-describing **in order to allow this.**

**Case22:** “in order to”没写成pattern

While a database system might choose to map each row to a contiguous storage container (e.g. a record) on a single disk page, it might also choose to store large, possibly infrequently referenced attributes of a table corresponding to large text objects, JPEG images, or multidimensional arrays in separate storage containers on different disk pages and/or different storage volumes **in order to maximize the overall performance of the system.**

**Case35:** “aim at” 没写成pattern

In addition, data analysis using data cubes has made huge advances, and now efforts are focused on integrating machine learning algorithms that infer trends, do data clustering, and detect anomalies. **All these tools are aimed at making it easy to analyze commercial data, but they are equally applicable to scientific data analysis.**

**---------------------------------------------------------------------------------------------------------------**

**Case25:** 句子被其他pattern优先匹配了 p\_2

Views serve many purposes including increased security (by hiding attributes from applications and/or users without a legitimate need for access) and enhanced performance (by materializing views defined by complex SQL queries over very large input tables).

这里第一处by被匹配后成功得到了原因片段(by hiding attributes …for access), 后面的片段就被抛弃了，没有进行匹配

**Case48:** pattern用的不对 p\_2

**Because data is so large, and IO bandwidth is not keeping pace,** moving code to data will be essential to performance.

应当用because的pattern，却被so优先匹配了，导致原因结果片段不对

**------------------------------------------------------**

**Case15:** Pattern写的太死 p\_3

**This self-description, or metadata, is central to all these scenarios;** it enables generic tools to understand the data, and it enables people to understand the data.

Enable要求前面匹配的代词只能是 this that these those，不能匹配it

**Case19:** Pattern 写的太死 p\_3

**In the next decade, as data interchange among scientific disciplines becomes increasingly important**, a common HDF-like format and package for all the sciences will likely emerge.

As的pattern不能处理这种句型，只能处理 &R as &C的句型

**Case31:** Pattern 写的太死 p\_3

**As file systems grow to petabyte-scale archives with billions of files,** the science community must create a synthesis of database systems and file systems.

As的pattern不能处理这种句型，只能处理 &R as &C的句型

**Case46:** Pattern写的太死 p\_3

Good metadata becomes central **for data sharing among different disciplines and for data analysis and visualization tools.**

For的pattern不能处理后接名词短语的句型

**------------------------------------------------------**

**Case40:** Stanford-parser解析句子有误p\_4

**Once you can put your types and your programs inside the database** you get the parallelism, non-procedural query, and data independence advantages of traditional database systems.

Stanford-parser将 “you get the paralleism” 归入了inside the database作为介词短语，而不是作为一个新句子的主谓。

**---------------------------------------------------------------------------------------------------------------**

**Case5:** 多句q\_1

The demand for tools and computational resources to perform scientific data-analysis is rising even faster than data volumes. **This is a consequence of three phenomena: (1) More sophisticated algorithms consume more instructions to analyze each byte. (2) Many analysis algorithms are super-linear, often needing N2 or N3 time to process N data points. And (3) IO bandwidth has not kept pace with storage capacity.**

**Case41:** 多句q\_1

**They have a schema language (metadata) to define the metadata. They have a few indexing strategies, and a simple data manipulation language. They have the start of non-procedural and parallel programming. And, they have a collection of tools to create, access, search, and visualize the data.** So, in our view they are simple database systems.

**Case43:** 多句q\_1



**------------------------------------------------------**

**Case4:** Implicit cause-effect link q\_2

Things were simpler then; **one could focus on the science rather than needing to be an information-technology-professional with expertise in arcane computer data analysis tools.**

**Case18:** Implicit cause-effect link q\_2

**Ideally much of this metadata would be automatically generated and managed as part of the workflow,** reducing the scientist’s intellectual burden.

**Case32:** Implicit cause-effect link q\_2

Set-oriented file processing will make file names increasingly irrelevant – **analysis will be applied to “all data with these attributes” rather than working on a list of file/directory names or name patterns.**

**Case33:** Implicit cause-effect link q\_2

Indeed, the files themselves may become irrelevant **(they are just containers for data.)**

**Case37:** Implicit cause-effect link q\_2

**This collection of problems is generally called the impedance mismatch – meaning the mismatch between the programming model and the database capabilities.** The impedance mismatch has made it difficult to map many science applications into conventional tabular database systems.

**Case39:** Implicit cause-effect link q\_2

Queries can read and write these extended types **using the same techniques they use on native types.**

**------------------------------------------------------**

**Case8:** Implicit cause-effect link q\_2 多句q\_1

**The new work style in these scientific domains is to send questions to applications running at a data center and get back answers, rather than to bulk-copy raw data from the archive to your local server for further analysis. Indeed, there is an emerging trend to store a personal workspace (a MyDB) at the data center and deposit answers there.** This minimizes data movement and allows collaboration among a group of scientists doing joint analysis. These personal workspaces are also a vehicle for data analysis groups to collaborate.

**Case13:** Implicit cause-effect link q\_2 多句q\_1

**Peta-scale data sets will require 1000-10,000 disks and thousands of compute nodes. At any one time some of the disks and some of the nodes will be broken.** Such systems have to have a mechanism in place to protect against data loss, and provide availability even with a less than full configuration — a self-healing system is required

**---------------------------------------------------------------------------------------------------------------**

统计未召回例子

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 未召回 | pattern没写 | pattern匹配问题 | | | | 不符合pattern匹配要求 | | |
| p\_1 | p\_2 | p\_3 | p\_4 | q\_1 | q\_2 | q1+q2 |
| 48 | 24 | 6 | 0 | 2 | 4 | 1 | 3 | 6 | 2 |

## F0029

人工标记36，共24例完全召回

**---------------------------------------------------------------------------------------------------------------**

**Case14:** ‘when’没写成pattern

When the objectID1-run2 pair is a miss, then objectID2 is zero, and the hitOrMiss flag suggests why (Ephemeral, Masked, or Edge)

**Case30:** ‘when’没写成pattern

**When considering SDSS observations separated by days or years,** only very slow-moving objects can be detected with the cross-match techniques here.

**Case21:** “caused by”没写成pattern

Matches are not transitive. For example, in Figure 2 object O1 matches O2 and O2 matches O3 but object O1 may not match O3. This might be caused by the object moving, or it might just be an unusually large position error, or they might just be different objects.

**Case31:** “given”没写成pattern

**Given this rather low object surface density (when compared to the Galactic Plane or the Galactic Center),** the techniques described here can find slowly moving objects by using a larger classification distance.

**Case36:** “given”没写成pattern

**Given the bundle and match tables,** the pivoted table can be constructed, using zero rather than null for missing objIDs, as follows:

**---------------------------------------------------------------------------------------------------------------**

**Case16:** 句子优先被其他pattern匹配了p\_2

**Indeed, the SDSS catalog pre-computes the spatial join as the Neighbors table using the Zones algorithm described in [2].** So the hit query is even simpler -- one just looks for neighbors within 1 “with run1≠run2 since Neighbors stores all object pairs within 30” .

这个句子优先被 ‘since’的pattern匹配了，结果片段 “so … run2”不能与前一句凑一起被so的pattern匹配

**----------------------------------------------------**

**Case24:** pattern写的太死了p\_3

It is easy to compute the aggregate statistics for the bundle table **once each match record has an assigned bundle ID.**

Once的pattern只能匹配once &C，&R的句型

**---------------------------------------------------------------------------------------------------------------**

**Case10:** 多句q\_1



**---------------------------------------------------**

**Case1:** Implicit cause-effect link q\_2

**Given several observations of the sky, called runs, astronomers often want to cross-match all the observations of each object from all runs that observed that object.** A typical first step is to process the runs to make an object catalog.

**Case4:** Implicit cause-effect link q\_2

**The observational errors, motions, and sizes all create their own errors, which must be convolved with this distribution.** These convolutions will broaden the Dirac delta.

**Case18:** Implicit cause-effect link q\_2

**This representation dovetails with the HTM library** [1] that makes it easy to find all points within a region.

**Case22:** Implicit cause-effect link q\_2

**Having the Match table makes it easy to reason about the observations of the same object and easy to collect statistics (average, variance,…) about the object’s position, magnitude, classification, the number and types of misses that the object experienced, and other attributes.** This suggests creating a Bundle table that records these statistics.

**---------------------------------------------------------------------------------------------------------------**

统计未召回例子

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Total | 未召回 | pattern没写 | pattern匹配问题 | | | | 不符合pattern匹配要求 | | |
| p\_1 | p\_2 | p\_3 | p\_4 | q\_1 | q\_2 | q1+q2 |
| 36 | 12 | 5 | 0 | 1 | 1 | 0 | 1 | 4 | 0 |