

TDT 2004: Annotation Manual

Version 1.2 – August 4, 2004

<http://www ldc.upenn.edu/Projects/TDT2004>

Table of Contents

1. Overview.....	3
2. Terms and Concepts	4
2.1 Topics, events and activities.....	4
2.2 Rules of Interpretation	5
3. Topic Selection	7
3.1 Strategy	7
3.2 Topic Profile.....	8
4. Topic Research.....	10
5. Topic Labeling	11
5.1 Overview.....	11
5.2 Relevance Labels	11
5.3 Search Guided Annotation Process.....	12
5.3.1 Stage One	12
5.3.2 Stage Two	12
5.3.3 Stage Three.....	13
5.3.4 Stage Four.....	13
5.4 Completing Annotation	13
6. Quality Control	14
6.1 Precision.....	14
6.2 Adjudication	14

1. OVERVIEW

In the Topic Detection and Tracking (TDT) project, researchers build algorithms for discovering and threading together topically related material in streams of news data in English, Chinese and Arabic.

TDT 2004 is the seventh in a series of open technology evaluations that cover several research areas: high accuracy retrieval of documents, text filtering, cross-language issues, machine translation, text segmentation, compensating for degraded quality text, novelty detection and other similar topics. In support of TDT 2004, the LDC will prepare a new corpus called TDT-5.

The data for the TDT-5 corpus was newly collected from English, Chinese and Arabic from April-September 2003. Unlike previous TDT corpora, TDT-5 does not contain any broadcast news data; all sources are newswire.

TDT-5 Corpus Content		
Language	Source	Doc Count
Arabic	AFA (Agence France Presse)	30,593
Arabic	ANN (An-Nahar)	8162
Arabic	UMM (Ummah)	1104
Arabic	XIA (Xinhua)	33,051
Arabic	<i>Total</i>	<i>72,910</i>
English	AFE (Agence France Presse)	95,432
English	APE (Associated Press)	104,941
English	CNE (CNN)	1117
English	LAT (LA Times/Washington Post)	6692
English	NYT (New York Times)	12,024
English	UME (Ummah)	1101
English	XIE (Xinhua)	56,802
English	<i>Total</i>	<i>278,109</i>
Mandarin	AFC (Agence France Presse)	5655
Mandarin	CNA (China News Agency)	4569
Mandarin	XIN (Xinhua)	37,251
Mandarin	ZBN (Zaobao News)	9011
Mandarin	<i>Total</i>	<i>56,486</i>
Corpus	Total	407,505

Table 1: TDT-5 Corpus Content

The corpus will include both original documents and English translations of the Chinese and Arabic data.

Annotation tasks for TDT 2004 include:

- **topic selection:** choosing 250 event-based topics across 3 languages
- **topic definition:** describing each topic's "seminal event", along with basic factual information
- **topic research:** researching and describing each topic's larger context (background, key players, terminology, related issues)
- **topic labeling:** finding the documents in the TDT-5 corpus that discuss the topic
- **quality control:** checking the results of topic labeling for completeness and accuracy
- **adjudication:** comparing LDC's human annotations with automatic system results

2. TERMS AND CONCEPTS

2.1 Topics, events and activities

Topic? Event? Activity? One of the biggest points of confusion for new annotators in TDT is the notion of *topic* and *event*.

A TDT **event** is defined as a particular thing that happens at a specific time and place, along with all necessary preconditions and unavoidable consequences. A TDT event might be a particular plane crash, or a single meeting, or a particular court hearing. An **activity** is a connected set of events that have a common focus or purpose, happening at a specific place and time; for instance, a campaign, or an investigation, or a disaster relief effort. For the purposes of TDT, a **topic** is defined as **an event or activity, along with all directly related events and activities**.

An example:

In February 1998, a low-flying U.S. Marine jet sliced through the cable supporting a funicular at a ski resort in Cavelese, Italy. The funicular then came crashing down, killing 20 people and injuring many more. The funicular's fall to the ground and the subsequent deaths and injuries were all unavoidable consequences of the jet flying into the cable, and are thus considered part of the same *seminal* event.

To get from **event** to **topic**, we have to consider what other events are directly related to the seminal event. In the case of the cable car crash described above, the rescue efforts, the victims' funerals, statements made by the US Marines

about policies for training missions in civilian areas, and the criminal investigation that followed the accident are all directly related to the seminal event. Therefore, all of these related events are part of the same TDT topic.

But what about things that are further removed from the seminal event? For instance, in this example, the jet that caused the accident was part of the NATO-led force patrolling skies over Bosnia during the war. It is possible to draw a connection between the war in Bosnia, the jet flying on this particular training mission, and the accident at the ski resort. Does that mean that the war in Bosnia is part of the same topic? **NO!** It's important to understand that TDT topics consist of a seminal event plus any **directly related** events and activities.

2.2 Rules of Interpretation

To increase the consistency of judgments about what constitutes a directly related event or activity, annotators refer to a set of ***rules of interpretation***. These rules state, for each type of seminal event – crimes, natural disasters, scientific discoveries, accidents, scandals, etc. – what other kinds of events are directly related. This then informs the annotators' judgments about the scope of the topic, and about which documents should be labeled "on-topic".

The thirteen types of seminal events, with examples of each drawn from the TDT-3 Y2001 Training Topics, are:

1. **Elections**, e.g. 30030: Taipei Mayoral Elections
Seminal events include: a specific political campaign, election day coverage, inauguration, voter turnouts, election results, protests, reaction.
Topic includes: the entire election process, from announcements of a candidate's intention to run through the campaign, nominations, election process and through the inauguration
2. **Scandals/Hearings**, e.g. 30038: Olympic Bribery Scandal
Seminal events include: media coverage of a particular scandal or hearing, evidence gathering, investigations, legal proceedings, hearings, public opinion coverage.
Topic includes: everything from the initial coverage of the scandal through the investigation and resolution.
3. **Legal/Criminal Cases**, e.g. 30003: Pinochet Trial
Seminal events include: the crime itself, arrests, investigations, legal proceedings, verdicts and sentencing.
Topic includes: the entire process from the coverage of the initial crime through the entire investigation, trial and outcome. Changes in laws/policies as a result of a crime are not generally on-topic unless a clear and direct connection between the specific crime and the legislation is made.
4. **Natural Disasters**, e.g., 30002: Hurricane Mitch
Seminal events include: weather events (El Nino, tornadoes, hurricanes, floods, droughts), other natural events like volcanic eruptions, wildfires, famines and the like, rescue efforts, coverage of economic or human impact of the disaster.
Topic includes: the causal (weather/natural) activity including predictions thereof, the disaster itself, victims and other losses, evacuations and rescue/relief efforts.

5. **Accidents**, e.g., 30014: Nigerian Gas Line Fire

Seminal events include: transportation disasters, building fires, explosions and the like.

Topic includes: causal activities and all their unavoidable consequences like death tolls, injuries, economic losses, investigations and any legal proceedings, victims' efforts for compensation.

6. **Acts of Violence or War**, e.g., 30034: Indonesia/East Timor Conflict

Seminal events include: a specific act of violence or terrorism or series of directly related incidents (such as a strike and retaliation).

Topic includes: Direct causes and consequences of a particular act of violence such as preparations (including technological/weapons development), coverage of the particular action, casualties/loss of life, negotiations to resolve the conflict, direct consequences including retaliatory strikes. This topic type is difficult to define across the board, and can easily become extremely broad and far-reaching. As such, each topic of this type is treated individually and is defined in such a way as to sensibly limit its scope and make annotation manageable.

7. **Science and Discovery News**, e.g., 31019: AIDS Vaccine Testing Begins

Seminal events include: announcement of a discovery or breakthrough, technological advances, awards or recognition of a scientific achievement.

Topic includes: Any aspect of the discovery, impact on everyday life, the researchers or scientists involved, descriptions of research and technology directly involved in the discovery.

8. **Financial News**, e.g., 30033: Euro Introduced

Seminal events include: specific economic or financial announcements (like a specific merger or bankruptcy announcement); reactions to the event; direct impact on the economy or business world. General economic trends or patterns without a clear seminal event are not appropriate as TDT topics.

Topic includes: the specific event, its direct causes, impacts on finance, government interventions or investigations, public or business world reactions, media coverage and analysis of the event.

9. **New Laws**, e.g., 30009: Anti-Doping Proposals

Seminal events include: announcement of new legislation or proposals, acceptance or denial of the legislation, reactions.

Topic includes: the entire process, from announcement of the proposal, lobbying or campaigning, voting surrounding the legislation, reactions from within the political world and from the public, challenges to the proposal, analysis and opinion pieces concerning the legislation.

10. **Sports News**, e.g., 31016: ATP Tennis Tournament

Seminal events include: a particular sporting event or tournament, sports awards, coverage of a particular athlete's injury, retirement or the like.

Topic includes: training or preparations for a competition, the game itself, results. For tournament and championship events like the World Series or SuperBowl, only direct precedents are considered on topic. Therefore, semi-finals and finals games leading up to the championship are on topic, but regular season play is not.

11. **Political and Diplomatic Meetings**, e.g., 30018: Tony Blair Visits China

Seminal events include: preparations for the meeting, the meeting itself, decisions, outcomes, reactions.

Topic includes: the whole process from the preparations and travel, the meeting itself, media coverage and public reaction, any outcome including legislation or

policies adopted as a *direct* outcome of the meeting. Sources often report on one of a series of meetings between two officials or delegations; in these cases, only the *current* meeting is part of the topic, although planning for a future meeting that is a direct outcome of the current meeting and is discussed as part of the current meeting will be considered on topic.

12. **Celebrity and Human Interest News**, e.g., [31036: Joe DiMaggio Illness](#)
Seminal events include: most often involves the death of a famous person or other significant life events like marriage.

Topic includes the specific event, causes (such as illness in the case of a celebrity's death) or consequences (such as a funeral or memorial service), public reaction or media coverage, editorials and opinion pieces, retrospectives or life histories that are a direct consequence of the seminal event.

13. **Miscellaneous News**, e.g., [31024: South Africa to Buy \\$5 Billion in Weapons](#)
Seminal events include all specific events or activities that do not fall into one of the above categories.

Topic includes the event itself, direct causes and unavoidable consequences thereof.

It is important to highlight the difference between a TDT topic and the notion of topic in normal discourse. While one might normally think of a topic as something broad like "accidents", a TDT topic is limited to a specific accident, like the cable car crash.

This particular conceptualization of topic is a critical component of TDT annotation, as it allows annotators to potentially identify *all* the stories in the corpus that discuss some pre-defined topic. The topic definitions and rules of interpretation ensure that each annotator is working with the same understanding of the topic at hand and, at least in theory, that all annotators will identify the same stories as on-topic. With these basic concepts as a foundation, TDT annotators select, define and research topics prior to beginning topic labeling.

3. TOPIC SELECTION

In TDT 2004, annotators will select, define, research and annotate a total of 250 topics.

3.1 Strategy

The topics are selected from a stratified, random sample of documents drawn from the fifteen English, Chinese and Arabic newswire sources collected from April through September 2003. The stratified random selection method gives each month of data from each source an equal chance of contributing a topic, although no effort is made to ensure equal representation of each source/month in the final set of selected topics. Within any month of data from a source, seed stories are selected at random. Annotators review randomized lists of 100 seed stories for each language.

A good topic is one that has a seminal event; that is, a particular thing that happens at a specific place and time. Unlike previous TDT topics, there are **no restrictions** on topic size, granularity or potential overlap. Good topics can be singletons (discussed in only one document in the corpus), small monolingual topics (covered in only one source or one language) or large multilingual topics (with broad coverage in all three languages). They can be very fine-grained (e.g., someone wins the lottery) or coarse-grained (e.g., an ongoing political scandal with lots of repercussions). They can occur at a single point in time or over a longer time span. It's also possible for two or more topics to overlap. For instance, there was a big power outage in Canada and the US on August 14, 2003. It would be perfectly acceptable to have one topic covering the outage in Southern Canada, and another topic focusing on the outage in New York City.

3.2 Topic Profile

Annotators also complete a basic profile for each topic they select. The topic profile consists of:

- **Topic title:** a brief phrase (under 10 words) that is easy to remember and immediately evokes the topic
- **Seminal event:** a brief description of 1-2 sentences; this should flesh out the title
- **What:** a statement of what happened during the seminal event
- **Who:** who (person, organization) was involved in the seminal event
- **When:** when the seminal event occurred
- **Where:** where the seminal event occurred
- **Topic size:** annotator's estimate of topic size
- **Language profile:** annotator's estimate of whether the topic will be discussed in other languages

Annotators can also review the topic profile for topics that have already been selected for any language.

An initial set of 125 good topics will be selected for each language. Once this is complete, team leaders will review the entire set of topics, whittling it down to a total of 250, roughly equally divided among the three languages. Approximately 50-75 topics will also be designated as "multilingual" topics; these will be annotated in more than one language. The non-multilingual topics will be annotated for one language only.

3.4. Topic Definition

Once a topic has been selected for inclusion in the final set of 250, it is converted into a full-fledged TDT topic through the process of topic definition. Each topic definition contains information from the topic profile created during the topic selection process, plus a few extra elements:

- **Topic Number**
- **Topic Explication:** provides a prose description of the topic's content.
- **On Topic:** this section applies the rule of interpretation to the seminal event and spells out what other kinds of events should be included in the topic's scope.
- **Notes (optional):** warns the annotator of potentially confusing or difficult aspects of the topic, and may explicitly limit the scope of the topic for purposes of annotation.
- **Rule of Interpretation:** link to the related rule of interpretation for this topic
- **Seed Story:** link to the topic's seed story (with English translation, for multilingual topics)
- **Chinese, Arabic keywords (optional):** provided for multilingual topics
- **Topic research:** link to the topic research document containing additional information

A sample topic description document appears below.

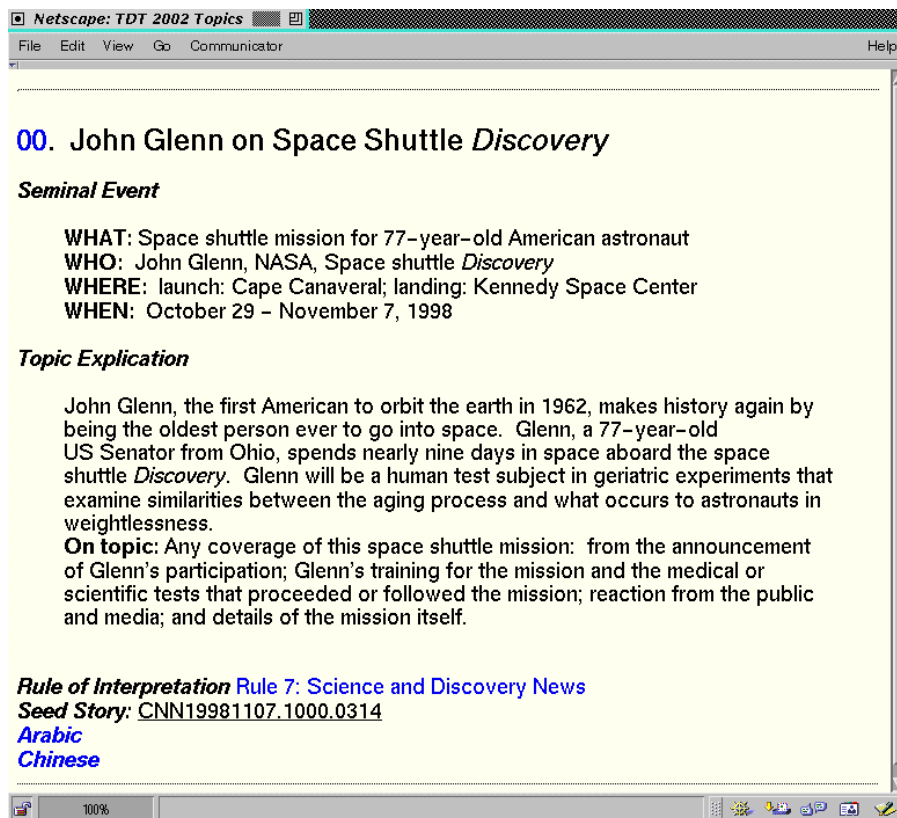



Figure 1. A sample topic definition

4. TOPIC RESEARCH

One of the biggest challenges for annotators is the task of keeping abreast of developments for a particular topic and understanding the scope of a topic in the corpus. Although the topic definitions spell out what sorts of stories might be considered on-topic, it is impossible to know in advance from having examined only one seed story how the topic might develop over time. In order to put the topics into a larger context, annotators conduct topic research, developing additional material like timelines, maps, keywords, named entities and links to non-TDT online resources for each topic. In addition, annotators use the topic research process to investigate how a given topic might be reported on across the sources, given the fact of media bias and the likelihood that different language sources might emphasize different aspects of the topic.

In search guided annotation, topic research feeds directly into topic labeling. Annotators submit parts of the topic research document as a query to the search engine during one stage of annotation. Topic research is a valuable resource not only for topic labeling, but also at later stages of quality control, when it provides a framework to monitor topic development and curb “topic drift”. Topic research is always accessible to annotators and is updated as the project and the topics evolve. A sample topic research page follows:

Topic 00
John Glenn on Space Shuttle 'Discovery'
Bethany Klein



We're not stopping any time soon, so if you need to use the bathroom before we take off please raise your hand now.

John Glenn, the first American to orbit the Earth in 1962, makes history again by being the oldest person ever to go into space. Glenn, at 77, spends nearly nine days in space aboard the space shuttle 'Discovery'. Glenn will be a human test subject in geriatric experiments that examine similarities between the aging process and what occurs to astronauts in weightlessness. Older people, for example, tend to lose bone and muscle mass, have trouble sleeping and experience decreased cardiovascular strength. That also happens to astronauts in space, but they soon recover on Earth.

Timeline

6.21.1997: Glenn volunteers for a space shuttle mission if NASA decides to study how weightlessness affects an older person.

1.16.1998: At NASA news conference, NASA administrator Daniel S. Goldin announces plans to send Glenn on the shuttle. Q and A session opens up concerning charges of political favoritism as a motivation for the decision. These questions, as well as questions regarding the validity and worth of Glenn's voyage, continue throughout the time period.

8.13.1998: Glenn takes a break from his Senate duties to be a full-time astronaut. From this point until the launch he trains almost continuously in Houston and Florida.

10.9.98: According to an ABCnews survey, nearly six in 10 say, whatever the reason (real scientific purposes or mainly as a p.r. gesture), it's good enough for Glenn to go on the mission.

10.29.98: Glenn and six other crew members meet with family members one last time before boarding the space shuttle discovery.

John Glenn is officially on his second trip orbiting the planet, a nine-day, zero-G journey that makes him the world's oldest astronaut. The afternoon launch was a success, despite a small panel falling off the shuttle Discovery during liftoff.

Glenn begins his geriatric research which includes sleeping nights in a cumbersome head net and body suit equipped with 23 sensors as well as being hooked up to a heart monitor with just seven electrodes and a miniature data recorder.

11.4.98: Glenn appears in an air-to-ground interview on Jay Leno's "The Tonight Show".

11.7.98: The shuttle crew members have a perfect landing at the Kennedy Space Center in Florida, capping a flight that returned the 77-year-old Glenn to space after 36 years. His heart and blood pressure were recorded during landing, and four hours of medical evaluation awaited him after he debarked.

11.8.98: Glenn goes to Johnson Space Center in Houston for three more weeks of additional medical tests. He and his fellow astronauts join family and friends there to celebrate their return.

Figure 2. A sample topic research document

5. TOPIC LABELING

5.1 Overview

Early stages of the TDT project employed a brute-force approach to topic labeling, in the sense that every story was read and exhaustively labeled against every topic. Beginning with the TDT-3 2000 Evaluation, a search guided annotation strategy was adopted. Previous experiments had shown that search guided annotation could produce results as good as brute force annotation while reducing costs and the effect of annotator fatigue. In search guided annotation, individual annotators work with one topic at a time. Whenever possible, the annotator who does the topic labeling is the same person who selected, defined and researched the topic.

In search guided labeling, the annotator uses LDC's EZQuery search engine to make multiple passes over the corpus, using different search strategies in an attempt to find all stories discussing the topic at hand.

5.2 Relevance Labels

Documents are labeled for their relevance to a topic using a two-way relevance scale, depending on their content:

YES: this story provides some amount of substantive information about the topic, no matter how little.

NO: this story does not discuss the topic at all, or only mentions the topic in passing without giving any information about the topic.

Although most **YES/NO** decisions are relatively straightforward, some decisions are difficult. Annotators are instructed to treat difficult cases as follows:

If you're having trouble deciding between YES and NO, ask yourself whether you learned anything about the topic by reading the story, no matter how small and no matter if you've seen that same information before. If you learn something about the topic by reading the story, then it should count as YES. If you're still having trouble making up your mind, consult with your team leaders and post a message to the TDT mailer. When in doubt, a story will usually fall on the side of YES.

Annotators may also choose to label a story with an optional **NOT EASY** label. All documents must receive either a **YES** or **NO** relevance label, but if an annotator struggles with a particular relevance judgment, s/he may add the additional **NOT EASY** label, which alerts team leaders of particular difficulties within a topic and triggers additional post-annotation quality control measures.

5.3 Search Guided Annotation Process

Search guided annotation implements four distinct stages, all of which utilize the EZQuery search engine. In each stage, annotators submit a different kind of query then label the relevance ranked list of documents that is returned. For each stage the list is limited to 200 documents.

Annotators continue working on one stage of annotation until the time allocated for that stage has expired, or until they have reached the "off topic threshold", whichever comes first.

Time limits for each stage of annotation are as follows:

Stage	Description	Time Limit
One	Search on seed story	60 minutes
Two	Search on topic profile	45 minutes
Three	Search on all YES stories	45 minutes
Four	Creative searching	30 minutes

A timer in the annotation tool warns users when they are approaching their time limit for a given stage.

The off topic threshold is defined as a 2:1 ratio of off topic to on topic stories, provided that at least the last 10 stories in a row are marked off topic. For instance, if an annotator finds 10 on topic stories in their list, they must also label at least 20 off topic stories, and at least the last 10 on the list must be labeled off topic before they can move on to the next stage of annotation. If the annotator finds no on topic stories on their list, they must label at least the first 50 documents before moving on to the next stage of annotation.

The annotation tool keeps track of the off topic threshold, and warns users when they try to move to a new stage prematurely.

5.3.1 Stage One

Stage one involves submitting the initial seed story as a search query. The EZQuery search engine returns a relevance ranked list of 200 documents and the annotator reads and labels each of these documents as **YES** or **NO**. Annotators work on Stage One for up to one hour, or until they reach the off topic threshold.

5.3.2 Stage Two

During Stage Two, annotators issue new queries using text selected from the topic documentation, which includes the topic profile and topic research

documents. Because the topic documentation is in English, annotators translate portions of the documentation for non-English topics.

Annotators work on Stage Two for up to 45 minutes, or until they reach the off topic threshold.

5.3.3 Stage Three

In Stage Three, annotators submit on topic stories identified in one of the first two stages as a query to the search engine. Annotators can review the previously labeled on topic stories and choose which ones to submit as a query, though the default choice is to submit **all** on topic stories.

Annotators work on Stage Three for up to 45 minutes, or until they reach the off topic threshold.

5.3.4 Stage Four

In Stage Four, annotators are encouraged to think creatively. By this point they have worked on the topic for some time. They have become topic experts and are well positioned to draw on their specialized knowledge to find the remaining on topic documents that have not yet been identified. Additional searches during Stage Four might be based on keywords, names, particular on-topic stories, etc. Annotators might limit their search to stories that occur before, after, or within particular dates or they might choose to focus on a particular source.

Annotators are permitted to execute a total of three searches during Stage Four. They may work on Stage Four for up to 30 minutes, or until they reach the off topic threshold.

5.4 Completing Annotation

Annotators are required to execute multiple searches for each topic, using each of the strategies described above. Each stage of annotation is timed, and annotators will spend no more than three hours total labeling any given topic. Annotators should spend the most time on those stages of annotation that are producing the best results for a given topic.

Annotation for a topic is complete when an annotator believes that s/he has found all the relevant documents for that topic within the corpus. In order to include all 250 topics (past evaluations have had approximately 50 topics), annotation for TDT2004 will be **incomplete**. That is, there will be no guarantee that every story on each topic will have been located. Instead, a topic will be frozen after three hours. At that point, the annotator will indicate whether the topic is complete (s/he believes all on-topic stories have been identified) or

"incomplete" (s/he believes there are additional on-topic documents in the corpus that have not yet been located).

Team leaders observe annotation progress and ensure that an appropriate searches have been conducted before confirming that a topic is complete, but the annotator, who by now is an expert on the topic at hand, is ultimately the best judge of a topic's completion point.

6. QUALITY CONTROL

The quality control measures for search guided annotation are multiple, including a precision check, adjudication of sites' results, and dual annotation with discrepancy resolution.

6.1 Precision

During Precision QC, senior annotators review all stories labeled as **YES** to identify possible false alarms, stories erroneously labeled as on topic. Working with a modified version of the labeling interface and examining one topic at a time, senior annotators read each story and either verify it as on topic, or change it to **NO**. When possible, the precision check is performed by the same senior annotator who conducted topic research for that topic. During the precision check, annotators keep a sharp eye out for cases of "topic drift", when the definition of the topic varies across annotators, by language, or over the course of topic labeling. By referring back to the topic explication and rules of interpretation, topic research documentation and annotator e-mail archives discussing the topic, senior annotators exclude stories outside the scope of the topic. Team leaders independently verify all changes resulting from the precision check. Along with reviewing all **YES** documents, senior annotators also review all judgments with the "Not Easy" label during the precision check.

6.2 Adjudication

In order to identify misses (on topic stories that are not identified as such), LDC relies on adjudication of research sites' results. NIST provides LDC with each research site's results for the topic tracking task. The sites' systems are scored against the LDC's human-produced topic relevance tables, with the annotators' judgments taken as ground truth. Each system false alarm is a potential LDC miss. It is not feasible to completely adjudicate all cases where LDC annotators differ from system performance; the effort needed to adjudicate all the cases of discrepancy would exceed the original corpus creation effort¹. Instead, the LDC reviews cases where a majority of systems disagree with the original annotation and modifies the topic labels as required. In previous TDT corpus adjudication

¹ For example, in the case of the TDT-3 1999 Evaluation topics, NIST delivered results containing approximately 1.5 million topic-story tuples from 7 research sites.

efforts, the probability of a system false alarm correlating to an annotator miss grew in proportion to the number of systems reporting disagreement with the original annotation.