

# **High Accuracy Retrieval from Documents (HARD)**

**Annotation Guidelines, version 1.3**

**October 11, 2004**

## 1 Introduction

High Accuracy Retrieval from Documents (HARD) is an information retrieval project, the goal of which is to retrieve highly accurate answers to queries by leveraging additional information about the searcher and/or the search context. This goal is met by using techniques like passage retrieval and targeted interaction between researchers and query issuers. The HARD evaluation of 2004 is run as an evaluation track within TREC, the Text REtrieval Conference sponsored by the National Institute of Standards and Technology (NIST).

Linguistic Data Consortium's participation in the HARD 2004 corpus creation effort centers around three major annotation tasks: Topic Creation, Clarification Form completion, and Relevance Assessment. This document will describe each task in detail, and will explain LDC's annotation approach to each one. The corpus from which information will be retrieved is a collection of English newswire documents, and is described in Section 2, below.

The technological goal of HARD is to improve information retrieval systems like search engines. While a number of effective search engines exist, the technology behind HARD will distinguish itself by basing search results on a searcher's "profile." The research community is investigating an assortment of variables that will help them develop increasingly accurate retrieval systems. Among the areas of interest to researchers are: a searcher's search experience (whether or not that person is a novice at searching, the kinds of searches this person has conducted in the past), specific areas of interest, regional focus, knowledge of the topic, etc. For the 2004 HARD evaluation, we will highlight a fairly small set of values; however, one could imagine an expansive group of categories that could be used in the future.

## 2 Corpus

The 2004 HARD evaluation uses the HARD 2004 English newswire corpus, which was collected and distributed by Linguistic Data Consortium for the HARD project. This corpus encompasses one year (2003) of newswire data, from eight (8) sources: AFE (Agence France Presse – English), APE (Associated Press Newswire), CNE (Central News Agency Taiwan), LAT (Los Angeles Times), NYT (New York Times), SLN (Salon.com), UME (Ummah Press – English), and XIE (Xinhua News Agency – English).

This chart shows the document and token count for each source.

Newswire Source	Stories	Total number of Tokens	Average number of Tokens/Story
AFE	226,777	71,831,282	317
APE	236,735	93,294,590	394
CNE	3,674	797,194	217
LAT	34,145	16,260,698	476
NYT	27,835	16,673,040	599
SLN	3,070	4,710,495	1,534
UME	2,557	782,064	306
XIE	117,516	24,016,722	204
<b>TOTAL</b>	<b>652,309</b>	<b>228,366,085</b>	<b>AVERAGE: 506</b>

**NOTE:** An English "token" is a word. Therefore, the average number of tokens per story is the average number of words per story.

Linguistic Data Consortium harvested each of the sources and cleaned and standardized the documents, in house. “Cleaning” in this case refers to a series of automated processes that eliminate “noise” in documents. Additionally, each story is given a unique document ID that is based on a three-letter newswire source abbreviation, followed by the year, month, day and chronological sequence of publication. All headers and document IDs are standardized across sources as a part of the corpus collection process.

### 3 Topic Creation

We will create twenty (20) training topics and fifty (50) original evaluation topics for the HARD 2004 track. The topics will derive from annotators’ areas of interest within the corpus epoch, 2003. Since HARD is a track that runs within TREC, LDC uses the established TREC approach to the first part of topic development, providing a short title, a sentence-long query, and a paragraph-long narrative, each of which describes the topic in increasing detail. However, the second part of topic creation is unique to HARD, and is explained in section 3.1, Topic Metadata.

Annotators research the time period on the Internet or in encyclopedias, and invent a series of queries about subjects that interest them. To create potential topics, annotators use a web-based questionnaire, and input information gleaned from personal research. Team leaders select the best of the topics for the training and evaluation datasets. **All topic research must derive from sources not covered by the corpus.** This means that all primary topic investigation will be conducted in other sources than the eight listed in Section 2, above.

#### 3.1 Topic Metadata

In addition to supplying descriptive topic elements, HARD topic creators define a set of parameters that further restrict the queries. These parameters, called “Metadata,” include *Genre*, *Geography*, *Granularity*, *Familiarity*, *Subject*, and *Related Text*. Each category possesses a series of values or options that the topic creator selects while making his or her topic. The goal of the metadata is to develop a sort of personal profile that will differentiate users’ results. For example, an expert in a certain field of knowledge should retrieve very different results from a novice in that same field of knowledge.

These fields leverage additional information about the desired results. In addition to this metadata, for internal purposes, other data is collected from the query issuer—like age, sex, and profession. This personal profile information is confidential, and is not distributed to research sites.

**The following Metadata parameters relate to the status of an article as on-topic or off-topic.**

##### 3.1.1 Subject

*[Values: Arts, Commerce, Current Events, Entertainment, The Environment, Health & Medicine, History, Law, Leisure, Politics, Science, Society, Sports, or Technology]*

The *Subject* category will determine a document’s status as on- or off-topic. For example, if Annotator A were to choose “Sports” for her topic and found a document that addressed her topic from an “Arts” perspective, it would be off-topic, because it is not addressing her fundamental topic query. The *Subject* category is a disambiguation of the topic itself.

Although this parameter may be of assistance to annotators judging whether or not a document is on-topic, its primary purpose is to give LDC a means of categorizing and organizing queries, and to ensure that there is ample variety in the kinds of topics addressed in the queries.

##### 3.1.2 Genre

*[Values: News Article, Opinion/Editorial, Other, or Any]*

This parameter limits the style of articles returned. News articles will discuss the facts, but try not to render any judgment about them. Opinion/Editorial pieces may contain many facts, but the

primary emphasis will be to argue for a particular point of view. Occasionally, this line can become blurred, as it is difficult to find news pieces without an observable bias. It is up to each individual assessor to make this determination.

### 3.1.3 Geography

[Values: *US, non-US, or Any*]

The *Geography* parameter restricts the region discussed in the returned articles. Articles concerned with the state of affairs in other countries will not be welcome returns for queries in which the US value has been selected, even if the story comes from an American news source. On the other hand, an article from a non-US source which discusses only American affairs will fulfill the US value of this Metadata parameter. The country of origin does not play a factor in the assessment of relevance for the *Geography* parameter.

### 3.1.4 Familiarity

[Values: *1 (little) or 2 (much)*]

The *Familiarity* parameter describes the level of expertise of the query issuer within a particular topic area. If an annotator selects "little," the query should only return articles written for someone with no knowledge of the topic. The returns should not contain technical jargon or advanced concepts, unless they are meticulously explained. Likewise, queries which have been designated as "much" are expected to contain un-explained references to terms, characters, places, and concepts.

This is easily the most subjective of the Metadata parameters, and the most difficult to assess. As the corpus is entirely comprised of newswire articles, it may be difficult to find or identify the boundary between "little" and "much."

### 3.1.5 Granularity

[Values: *Document or Passage*]

This parameter concerns the level of resolution of the actual relevance assessment. Queries which have "Passage" selected for this parameter will receive a much higher-resolution assessment than those which have been designated for document-level analysis. The "Passage" search returns will be read in detail a second time, and the relevant sections of each document will be highlighted. This does not affect the amount of information desired in the returns, only the level of detail with which the returns will be read.

### 3.1.6 Related Text

[Values: *This parameter has no predetermined values*]

The *Related Text* parameter allows annotators to show the searchers two example documents, one which is relevant to the Metadata and is representative of the desired returns, and one which is on-topic but does not satisfy the Metadata parameters. **None of the *Related Text* examples will be drawn from the corpus (see Section 2).**

Although this parameter may help the researchers, it is also intended as an aid to the annotators. There is a significant time lapse between each phase of the HARD project, and this parameter gives annotators a chance to look back to the topic creation phase and re-experience their original visions for the queries that they initiated.

### 3.1.7 Metadata-Narrative

[Values: *This parameter has no predetermined values*]

This parameter allows annotators to indicate how they think the Metadata parameters that they chose will affect the results of the search. They should go through the list of Metadata parameters and indicate which they think will be the most constraining on the search returns. Like the *Related Text* parameter, the *Metadata-Narrative* is as much for the annotators as it is for the searchers. It serves as a justification for the chosen parameters, should an annotator be unable to remember why he had chosen particular values. It can also be used by the researchers to direct their results, and focus on the particular issues of the individual search.

## 3.2 How the Metadata affects relevance

The values of *Familiarity*, *Geography*, and *Genre* will affect a document's relevance to a topic. Each topic annotator starts with a fairly impersonal query or statement about a subject that interests her. Because the goal of HARD is to create personally-tailored responses to queries, not all documents that are relevant to the basic topic query will be fully relevant for the topic. Therefore, a necessary distinction is made between those documents that are **on-topic and satisfy all document-level metadata values (relevant)**, and those documents that are **on-topic but do not satisfy the metadata values (on-topic but not relevant)**.

Not all of the metadata, however, will affect relevance. The two *Related Text* regions, for example, are included as documentation aids to annotators, and provide relevance feedback cues for researchers. Similarly, the *Metadata-Narrative* text region is an opportunity for the topic creator to show why she chose the values she did.

## 4 Clarification Forms

Between the topic creation and topic assessment stages of HARD, research sites submit "clarification forms" to LDC for a round of relevance feedback. Clarification forms constitute an optional portion of the track for participating research groups. Those who choose to take part, submit HTML questionnaires to LDC in an effort to glean targeted information from the topic creator.

The clarification form stage occurs before the research sites receive each topic's metadata information. They generate HTML documents based solely on the topic statements, and may angle for more information in any way that is supported by both HTML and Netscape 4.78. Each LDC assessor completes the forms that pertain to his or her topics by viewing the HTML files on the Internet. After each of the 50 forms per site per run has been answered, LDC returns the results to the individual sites. **Annotators will spend no more than three (3) minutes completing each form.**

In 2003, we did not formalize a system for viewing and responding to the clarification forms. Instead, each individual assessor accessed the HTML directory for each site and responded to the forms in turn. This "system" is not preferable for two reasons: It is time-consuming; and annotators view all forms in the same order for every topic.

For HARD 2004, we will create an interface that will manage the forms for each topic. An assessor may need to judge up to 50 forms for each of her topics, so it will be necessary to impose an electronic system to maintain all of these files. Annotators will access the clarification form management tool, click on one of their topic numbers, and see a randomized list of clarification forms. They will read and respond to each form until they have completed all forms for that topic. Every time the annotator accesses this interface, the untouched forms for her topic will be randomized again. Randomizing the clarification order will reduce the influence of learning the styles of questions of each site, so that every site's forms are given equal weight in this process.

## 5 Relevance Assessment

### 5.1 Training Data

Our method for assessing relevance differs between training and evaluation datasets. In the first case, we will develop the 20 training topics in house, and search for our own documents. This approach is quite different from the evaluation set, where we will receive documents from researchers.

Each annotator submits the topic statement plus metadata to the corpus. A list of potentially relevant documents is produced by the search. Assessors read through the top 100 of these documents and determine their relevance to the topics. No passage-level assessment is performed for training topics.

## 5.2 Evaluation Data

### 5.2.1 Introduction to relevance assessment

After participating research sites test their retrieval systems on the fifty HARD topics, they send their results to NIST to be organized and scored. Sites' results are sets of documents that could be relevant to the topics. These document sets are retrieved in two separate ways. First, sites use only the topic descriptions to find relevant documents; next, they use the metadata information, in addition to any results gleaned from the clarification form process, to retrieve more accurate results. Both sets are sent to NIST. The results are pooled, and LDC is sent the top eighty-five (85) documents from each site for each topic for relevance assessment. With 85 documents per run the pools average 747.2 documents with a minimum of 352 documents and a maximum of 1442 documents.

### 5.2.2 Task Overview

Relevance assessment is the process where annotators check each document returned from the research sites to verify how relevant the document is to what the searcher was looking for. Ideally, the original topic creator will perform the relevance assessment for his or her topics. Using a tool created in-house for this task, each assessor will examine each document in the list, for the topics she created. Assessors review the topic description and all metadata selections before progressing to assessment. The topic description remains in front of the assessor throughout the process.

Assessors are not permitted to skip documents. After carefully reading each one, the annotator assigns one of three labels to it based on its relevance to the topic.

### 5.2.3 Labels Defined

LDC will use the following definitions for topic labeling:

**RELEVANT: YES (Y):** This article discusses the topic in a substantial way. Articles that annotators label YES should answer the topic query without a doubt. They do not have to contain new information about the topic; a story that summarizes a topic's history, or one that gives a snippet of information that has been read about before still counts as a YES. Even if the article contains only a sentence with on-topic, relevant information, it should be considered a YES.

**OFF-TOPIC: NO (N):** This article does not discuss the topic at all, mentions the topic in passing without giving any information about it, or fails to address the specific query of the topic. If an article simply names a topic, or makes reference to it, but does not provide any information about it, that article should be labeled NO.

**ON-TOPIC: METADATA (M):** This label is applied for cases where the article is relevant to the topic statement, but fails to meet the demands of the Metadata. An article might get labeled METADATA if it gives plenty of information and answers the topic query, but is an Opinion/Editorial piece and the parameter *Genre* is set to News Report.

Since stories that do not match the *Subject* parameter should be labeled NO, and *Granularity* is dealt with later, the only Metadata parameters that restrict content are *Genre*, *Geography*, and *Familiarity*. When an annotator chooses to label a document METADATA, he or she will also be asked to choose which of those three Metadata parameters is not satisfied by the document.

## 5.2.4 Making the decision

The decision between YES and NO stories is usually clear, but some cases may be difficult. If there is difficulty deciding between YES and NO, an assessor should determine whether they learned anything about the topic by reading the story and whether they can answer the topic query based on that information. If an assessor has a particularly hard time making a decision between YES and NO, or YES and METADATA, she should indicate that by checking the “*difficult decision*” box. Documents labeled difficult will receive closer attention from team leaders during quality control annotation.

## 5.3 Passage-Level Assessment

For the topics with a “Granularity” metadata field other than “Document” – that is, “Passage” – LDC will provide passage-level relevance judgments, as well. During the document-level relevance assessment task, all documents judged YES (relevant) are annotated for passage-level relevance.

Once the YES/*RELEVANT* or the METADATA/*ON-TOPIC* judgment is made, a separate annotation tool shows the text of that document. The assessor then selects the passages that are relevant and submits her selections. The smallest unit of a “passage” is two words; the largest unit, the entire document. Selection is performed by highlighting the passage with the computer mouse, and clicking “YES” or “Difficult decision.” The latter decision means only that the passage is still on-topic, but that the decision to select the passage was difficult. The selected passages appear in another window of the assessment tool. An assessor may review her selections again as a quality control measure, and may change her decisions at that point.