Query Relevance Guidelines
OpenCLIR Version

Document Date: 2019-02-13

1 Summary

In this project, annotators will decide whether a document is relevant to a search term or set of terms (for example, traditional medicine). The search term(s) are referred to as a "query"[1].

Queries will typically comprise 1 – 3 words, though they may contain more words in some cases. The format of the query will indicate the way in which the query should be understood. The relevance judgment will be a binary choice: the document(s) will be marked as either **relevant** or **not relevant** to the search term. All search term(s) will be in English. All documents will be in the "document language", that is, the original language of the document. Note that relevance is based on English queries and is carried out by annotators who are proficient in both English and the target language. It should be noted that any English words or phrases that occur in the documents relevant to the English query are deemed relevant.

2 Query formats

Queries will be presented in a number of formats. It is very important to understand clearly the differences between these formats. The formats are a guide to how the annotator should judge the relevance of the query to the document.

Broadly, there are three types of query**: simple**, **conceptual**, and **hybrid** queries.

2.1 Simple queries

[1] See https://www.nist.gov/sites/default/files/documents/2018/07/12/openclirqueryspecification.pdf for the complete specification of the MATERIAL query format.

For this query type, annotators will look for a *translation equivalent* in the document. If one or more translation equivalents are found, the document is relevant to the query.

A translation equivalent is defined as “a word or phrase in the document language that matches the semantic concept denoted by the query string”.

For example, a translation equivalent in Spanish for the simple English query phrase *my mother* would be *mi madre*, or alternatively, *mi mamá or madre mía.*

More than one query term may be included in a query. In this case, a translation equivalent for both terms must be present in the document. Note that if a query is meant to be interpreted as a phrase, it will be enclosed in quotation marks. E.g. *military, force* should be interpreted as two simple query terms: *military* and *force*. But *“military force”* must be interpreted as a single query term.

In languages where there is code-switching into English, the query term may be partly or wholly matched in English.

2.1.1 Examples of simple queries[2]

bribery

a single query term; a document is relevant if it contains at least one translation equivalent of the word *bribery*.

bribery, police

two query terms; a document is relevant if it contains at least one translation equivalent of the word *bribery* **and** at least one translation equivalent of the word *police*. Note that, unlike quoted multi-term queries below, the different terms may occur anywhere in the document and do not have to refer to the same entity or concept. For example, a document *“…police are investigating a robbery. … In another news, the mayor is accused of bribery…*” would be relevant to the query.

“military force”

a single query term; a document is relevant if it contains at least one translation equivalent of the phrase *“military force”.*

“military force”, attack

two query terms; a document is relevant if it contains at least one translation equivalent of the phrase *“military force”* **and** at least one translation equivalent of the word *attack*.

[2]

“paintings of (people on trains)” a single query term. Brackets may be provided within a multi-term query phrase to help avoid ambiguity in understanding the query. For example, the reading *“(paintings of people) on trains”* is excluded – that is, the people are on the trains, not the paintings.

nurse [hyp: profession] a single query term. Square brackets are used to clarify the sense of the word, where the word has more than one meaning in English. E.g. “nurse” in English may be a person in the medical profession of nursing, or it may mean “to breast-feed a baby”. Only translation equivalents for the profession can be considered relevant.

<contaminated>, zone two query terms; Angle brackets are used to limit the possible translation equivalents of the word. In this case, a document is relevant if it contains at least one **exact** translation equivalent of the word *contaminated* (i.e. **not** *contaminates*, *contaminating*, etc.) **and** at least one translation equivalent of the word *zone*.

2.1.2 Guidance for assessing relevance for simple queries

1. Grammatical forms

In general, a *translation equivalent* refers to any of the different grammatical forms of the word(s) in the document language, unless the query indicates otherwise. Whether the word is in the *past or present or future*, indicates a *continuous or not continuous* action, *one thing or more than one thing* (singular vs plural), is *negative or positive*, *formal or informal*, *male or female gender*, will **not affect the judgement of relevance**.

For example, if the query term is *computer* in English, then translation equivalents of *computer* and *computers* are both considered relevant. Angle brackets, <>, will be used to further constrain the query results to a specific grammatical form. When angle brackets are used with the query term, only a translation equivalent corresponding to the specific English grammatical form is to be considered relevant. For example, if the query term is <computers>, then only translation equivalents that specifically express a plural form will be considered relevant. In the example above, *computer* will not be considered relevant to the query.

2. Semantic scope

For a simple query, the scope of the word cannot be significantly different from the search term. For example, if the document contains a sub-part of the search term (e.g. “*forehead injury”* where the search term is “*head injury*”), then this will not be considered relevant. The same would be true for a term that is much broader than, and includes, the search term, e.g. *medicine* where the search term is *“herbal medicine”.*

In addition, when a particular object is mentioned in the query, there should be an actual reference to that object in the document. For example, the query *cat [hyp: animal]* suggests an actual cat. The English phrase “cat call” would not be relevant to this query, because there is no actual reference to the animal.

The table below presents some simple queries and sentences that are considered relevant/non-relevant based on the given query:

Simple Query	**Relevant Sentence**	**Non-relevant Sentence**
“head injury”	He sustained a head injury due to the incident.	He sustained a forehead injury due to the incident.
“herbal medicine”	Why don’t you try some herbal medicine?	Why don’t you try some medicine?
cat	The cat chased away the mice.	The Royal Navy used to use the cat-o’-nine-tails for punishment.
fruit [hyp: food]	The lemon tree has a lot of fruits.	They can finally enjoy the fruits of their labor.

3. Additional words

Additional words or inflectional morphology do not prevent a word or phrase in a document from being considered a translation equivalent, as long as the required words are also present.

For example, for the query “a pleasant journey”, the phrase in French *Un voyage très agréable* (literally, *a journey very pleasant*) would be considered relevant, despite the intervening word *très* (very) inside the phrase. Similarly, the phrase *Un voyage très agréable* would also be considered relevant to the query “a journey”, despite the French phrase containing an additional adjective and adverb.

To give another example, for the query “pet cat”, a document containing the phrase “cat:my pet” (e.g., Arabic قطتي الأليفة *qittaty al-alifah*, literally *cat:my the-domestic*) should be considered relevant, despite the addition of “my”, because the phrase contains a translation equivalent both for the word *cat* and the word *pet* (or *domestic*).

However, if a translation equivalent for either the word “cat” or “pet” are missing from the document, the document is not relevant, *even if the idea is implied*. This is an important way in which simple queries are different from conceptual queries, which are explained next.

For example, the French phrase *persan de compagnie* (Persian-companion (animal)) does not explicitly mention “cat”, though it is implied in the name of the breed (Persian). Therefore this phrase would not be considered relevant to the query “pet cat”.

2.2 Conceptual queries

For **conceptual** queries, annotators are expected to search for something in the document that is relevant to the **concept** or **topic** provided by the query term(s). Conceptual queries are indicated with a “+” symbol to show that annotators must search beyond the literal words to **include concepts** or **topics that are related to the search term.**

For example, for a query such as *beekeeping+*, the annotator would search the document for evidence of activity or material that is related to the practice of beekeeping. A relevant document might describe how to keep hives clean, for example. **The specific term *beekeeping* need not be present in the document for the document to be judged relevant**.

2.2.1 Examples of conceptual queries

beekeeping+ — a single query term. A document is relevant if it is “about” or mentions beekeeping and related practices, the history of beekeeping, the profession of beekeeping, challenges related to keeping bees, etc.

“freshwater fish”+ — a single query term. A document is relevant if it is “about” or mentions freshwater fish and their habitats generally, or specific types of freshwater fish (e.g. carp, bream, tilapia), etc.

EXAMPLE_OF(freshwater fish) a single query term (consisting of one or more words), but with the constraint that **only subtypes and instances of the target query are relevant to this query**; other topically related terms are not relevant.

“(social media) discussion”+ a single query term. Both “social media” and “discussion” would be open to interpretation. A conceptually relevant phrase in the document language might be, for example, the translated equivalent of “Twitter discussion” or “Facebook conversation”.

strike+ [evf: labor] a single query term that has been constrained to one of the English meanings of the concept “strike”. In this case, discussion of workers “on strike” due to pay conditions will be relevant; a mention of something being hit (“I will strike the target”), or of the way an idea comes to someone suddenly (“an idea struck him”) will NOT be relevant.

Note: More than one conceptual query term will not occur in the same query.

2.2.2 Guidance for assessing relevance of conceptual queries

1. Substantial discussion vs incidental mention

No distinction will be made between a substantial discussion of the concept or an incidental mention – any mention of the concept will suffice for the document to be relevant.

2. Grammatical forms

As for simple queries, relevant translations of a conceptual query refer to any of the different grammatical forms of the word(s), unless the query indicates otherwise. That is, whether the word is in the *past or present or future*, indicates a *continuous or not continuous* action, *one thing or more than one thing* (singular vs plural), is *negative or positive*, *formal or informal*, will **not affect the judgement of relevance**.

3. Supplementary knowledge

In some cases, (ontological) knowledge from structured online sources will be needed to identify relevant documents. For example, the query “freshwater fish”+ requires knowledge of the names of freshwater fish. To use English examples, one might need to know that "bream", "bass", "carp", "pike", "perch" and "trout" (and their equivalents in the document language) are words for fish, and that they are types of freshwater fish. Ontological knowledge of Proper Nouns will not be tested in the program. One need not know that “Peter O’Neill” is the Prime Minister of Papua New Guinea.

4. Complex topics

Some topics are more complex than others. In some cases, a number of different pieces of information may be needed to call the document relevant to the conceptual query topic.

For example, the query term *safari*+ suggests a complex set of ideas, including a location (usually African), a tour (usually guided) and the presence of wildlife (usually in their natural habitat). If only one or two of these ideas are present in a document, this may not be sufficient to make the document relevant to the query. If in doubt, annotators should check with their supervisor.

2.3 Hybrid queries

Hybrid queries combine a simple and a conceptual query term (or terms). **A document is only judged relevant if *both* search terms are relevant to the document**.

For example, the query *zoos+, environment* is a hybrid query. *zoos+* is the conceptual query part, as indicated by the “+” symbol, and environment is the simple query part. This query asks the annotator to search for (1) anything in the document that is substantially related to the term *zoos* as a concept (e.g. nature reserve, wildlife park, etc.) and (2) in addition to this, in order to be relevant, a document must also contain *a translation equivalent* of the simple query term *environment.*

2.3.1 Examples of hybrid queries

“traditional practices”, health+

two query terms: one simple (“*traditional practices”)* and one conceptual (*health+)*. A document is relevant if it is “about” or mentions health or health-related practices AND includes mention of a translation equivalent of *“traditional practices”* (traditional practice, customary practice, etc.)

EXAMPLE_OF(vaccination), immunity

two query terms: one simple (*immunity*) and one conceptual (*EXAMPLE_OF(vaccination)*). A document is relevant if it contains a translation equivalent of *immunity*, and also contains an example of a type of vaccination.

Appendix: Summary of query formats

x

x is any simple query

e.g. *dictator*

“x y”

x y is a single, multi-word term, and is to be treated as a simple query (as opposed to two separate terms subject to conjunction).

e.g. *“social media*”

x+

x is conceptual and is subject to semantic expansion.

e.g. *beekeeping+*

“x y”+

x y is a single, multi-word term, subject to semantic expansion.

e.g. *“climate change”+*

EXAMPLE_OF(x)

x is a single term (consisting of one or more words) subject to limited semantic expansion. Subtypes and instances of x are relevant to this query; other topically related terms are not relevant.

e.g. EXAMPLE_OF(freshwater fish); EXAMPLE_OF(hockey player)

“(x y) z”
“w (x y)”

x y is a constituent phrase within a larger phrase. Only one level of parentheses is allowed.

e.g. *“(social media) post”; “paintings of (people on trains)”*

x, y

x and y are separate terms subject to conjunction. The order of the two terms is unimportant. **We propose that conjunction of two conceptual terms be disallowed, at least in the base period**. Conjunction of one conceptual term and one simple term is allowed (and constitutes a hybrid query).

e.g. *oil, tax; vaccination+, autism; habitat, EXAMPLE_OF(rodent)*

x [t: y]

y is a constraint of type t on the query term x. Valid values for t are “hyp” (hypernym), “syn” (synonym), and “evf” (event frame).

e.g. *nurse [hyp: profession]; strike+ [evf: labor]; EXAMPLE_OF(virus [evf: medical]); retreat [syn: withdraw].*

<x>

x is a term that is subject to an English morphological constraint. Only simple (non-conceptual) terms may be morphologically constrained for a marked morphological property, e.g. <dogs> would require a plural mention to be relevant e.g. *<contaminated>, zone*

Note: disjunctive queries (e.g., x OR y) are disallowed.

Notes on scope

Constraints will always immediately follow the word or phrase they modify. If the constrained word or phrase is subject to conceptual expansion, the constraint follows the plus sign. We are permitting the following formats for constraints and operators.

“x y”+ [w]	The whole phrase “x y” is constrained and subject to conceptual expansion.[3]

“x [w] y”+	Only x is constrained, and the whole phrase is subject to conceptual expansion.

“x y [w]”+	Only y is constrained, and the whole phrase is subject to conceptual expansion.

Relevance of grammatical features

Unless the query is morphologically constrained, all inflectional forms of a query string are relevant,; e.g., tense, number, gender, polarity, etc.

[3] This combination is likely to be rare.