2008 NLP Shared Task

FAQ

Posted on 05-02-08

Q: The on-line Challenge documentation states "They provided judgments that were strictly based on text; we refer to these judgments as "textual judgments" and mark the source of these judgments as "textual" in the annotation files." It also states "Intuitive judgments are based on implicit information in the narrative text." It is not clear to us where to draw the line between "based on text" and "implicit information." Would, for example, a stated blood pressure that falls within the range considered to indicate hypertension be considered textual or implicit information? Furthermore, would a textual judgment imply an equivalent intuitive judgment? In other words, is the set of textual judgments of a given type a proper subset of the intuitive judgments of that type for a given disorder, or are the two types of judgments independent?

A: The line between textual and intuitive is hazy, as always. Doctors decided that if the judgment required any thinking, it was not explicit. For example, if they had to compute BMI to decide obesity, they marked that as intuitive. My understanding is stated blood pressure is explicit. Also, doctors' intuitive judgments very rarely disagree with explicitly stated information in text.

Q: Could you define the "questionable" category, and does the meaning depend on whether the context is textual or intuitive judgment? So for example, in the context of textual judgment, is a classification of "questionable " to be assigned to a document only if there is an explicit statement that the disorder is questionable, questioned, unconfirmed, etc. for that patient? Could the "questionable" classification also be appropriate based on medical reasoning applied to other information in the document in the textual judgment context, only in the intuitive judgment context, or in both contexts?

A: In the text, an unconfirmed statement such as "patient may have pneumonia" is enough to mark the judgment as questionable. This has to be an explicit expression of uncertainty. For intuitive judgments, if given the information in the document, the doctor is able to assert "the patient may have pneumonia" then the intuitive judgment would be questionable. Medical reasoning comes into play only in the case of intuitive judgments.

Q: Does temporality play a role? For example, if it is obvious from a report that the patient exhibited a co-morbidity in the past but does not have the condition at the moment, is it still marked up as "Yes" by the annotators?

A: (This answer is quoted directly from one of the annotators.) It may be easier to answer this question if there is a more specific example. If I intuitively thought the patient had a co-morbidity in the past and wasn't currently on a medication or if the disease was not listed, I would have entered an "intuitive - yes" if it was obvious or "Intuitive - Questionable".

Q: What where the instructions given to he annotators regarding the distinction between textual judgments and intuitive judgments? Are textual judgments based on the occurrence of the name of a condition or its synonyms in the text and no further information? For example, for obesity, if a report doesn't mention obesity or a synonym of obesity, but contains the height and weight of the patient from which it follows clearly that the patient is obese, would this be a textual "Unmentioned" and an intuitive "Yes" (even though the judgment is not really intuitive but based on clear data given in the text?). Also, for conditions whose presence can be inferred from quantitative measures, how is the presence of these conditions defined (e.g., above what BMI value is a patient considered obese?).

A: (Answer quoted directly from one of the annotators.) Good question. I kept it simple and based my textual answers strictly with words found in the text, i.e., obese, obesity...however if I intuitively thought the patient had obesity, which included calculating BMI, I marked this as "intuitive YES". I based obesity on current standards.

Q: Would it be possible to make the annotation guidelines that were used by the annotators available to the challenge participants?

A: The doctors were given instructions on how to distinguish between textual and intuitive judgments. The levels of certainty that would lead to "Y", "N", and "Q" judgments were discussed with them. They then refined the definitions they received. This information is posted on our web pages.

Posted on 04-17-08

Q: It is quite straightforward to interpret intuitive annotation as the physicians had to make a decision on textual unknown documents based on the expertise and impressions (Unmentioned to Yes/No/Questionable). Documents that caused disagreement were left out. There are some label-combinations that seem very weird, e.g.: Yes/No to Disagreement (left out, 212 docs/label pairs in total): this suggests that these documents hold textual evidence on the patient exhibiting a certain disease (or not exhibiting it) but in the intuitive judgment it caused disagreement. The probable disagreement is 1 Yes and 1 No decision as physicians seem to avoid using Questionable tag (infrequent in both textual and intuitive annotation). This emphasizes that 1 doctor denied the textual evidence that he probably found earlier /textual agreements are high/. How can this be interpreted? Have the physicians performed the two annotation in parallel, or if not, did they see the textual labels when performing intuitive annotation? There is 1 document that received textual No annotation and intuitive Yes annotation (Document 1072 for Congestive Heart Failure) that seems especially weird /textual evidence of not having chf and impression that he or she indeed

has chf/. Is it possibly an annotation error or if not, could it be interpreted somehow? Probably it would give a good insight to the logic of textual and intuitive judgments of the physicians.

A: Indeed, one doctor disagreed with the explicit information in the text in one case (Record #1072). But overall, there was agreement on the textual and intuitive judgments in order to have the record included in the final data. In other cases, the fact that there is tie-breaking for the textual but no tie-breaking for the intuitive judgments explains your observation. There are several categories of documents/judgments in the data. Assume three annotators: A, B, and C. A and B made the first pass annotations and C then broke textual judgment ties. So, the judgments you are looking at fall into the following categories:

1. A and B agreed on textual judgment and intuitive judgment (because they agreed on both)
2. A and C agreed on textual judgment while there is no intuitive judgment (because A and B disagreed on text and therefore on intuition)
3. B and C agreed on textual judgment while there is no intuitive judgment (because A and B disagreed on text and therefore on intuition)

Q: You say "The training set includes 60% of each of "Y", "N", "Q", and "U" textual judgments for obesity from the complete data set." Does this mean that there will be no new documents provided, just a template to make new judgments on the existing documents? Separately, will every textual judgment request be paired with an intuitive judgment request (and vice versa)?

A: The test documents won't be released until evaluation time. There is a set of test documents that will remain unseen until that time. Just like in the training set, there is no guarantee that there will be an intuitive judgment for each textual judgment in the test set.

Posted on 03-23-08

Q: How were the records de-identified?
A: The records were de-identified semi-automatically. An automatic and two manual passes were made over each record. The manual passes followed the automatic but were in parallel with each other. A third manual annotator resolved the disagreements that resulted from the two manual passes and finalized the identification of private health information. We then replaced the identified private health information in several ways. For names of doctors and patients, we drew random names from the US Census Bureau names dictionary. Therefore, the surrogate names in the records will look like real names but they do not belong to the actual patients. We made no effort to keep co-reference, i.e., for each occurrence of Dr. "John Smith", we drew another name from the US Census bureau dictionary. For phone numbers, ID numbers, and ages, we randomly generated surrogates by replacing each digit and each letter with a random digit and each letter with a random letter. For dates, we generated random dates as surrogates. For locations and hospital names, we randomly generated surrogates by permuting syllables of existing hospitals and locations. We complemented locations with legitimate city and state names, followed by random zip codes, where appropriate.

Posted on 03-19-08

Q: Documents 164 and 175 have no intuitive judgments. Was this intentional? What are the criteria for including a judgment?
A: Judgments reflect majority vote among annotators. For textual judgments, annotators were two obesity experts and a resident doctor. The obesity experts provided initial judgments and the resident doctor broke ties. The majority vote among them provided the final judgment. For intuitive judgments, only the obesity experts provided judgments. There was no tie breaking. The judgments reflect the agreements between the two experts.
Records 164 and 175 are indeed only textually annotated because one of the obesity experts did not get to annotate them. Given that those are the only two records in that position, we chose to eliminate them from the intuitive data set.

Q: Judgments for "Gout" are in the data set but the co-morbidity itself is not on the annotation sheet. Shall I ignore this co-morbidity?
A: It is now back on the annotation sheet. It was removed only accidentally.

Q: All the intuitive judgments for CHF are "Y". There are no "N"/"Q"/"U" and the total # of diseases is 243, much less than other co-morbidities. Is there anything missing?
A: Yes! Thanks for catching. Indeed, for CHF, only the "Y"'s were released. The rest are in the addendum file. Please remember to download this file, also.

Q: Of textual judgments, there are only 62 No's out of 2704 non-unmentioned judgments (2%), appearing in only 9 of 15 classes. Will the primary outcome measure for textual judgments be macro-averaged or micro-averaged across classes? (It appears to me that negation detection won't make a difference in the micro results.)
A: Please stay tuned for the exact evaluation formulae. It will be another week or two before we release them.

Q: Will you issue an official scoring script?
A: Yes! Please stay tuned!

Q: What do "Y", "N", "U", and "Q" used in the annotation files stand for?
A:"Y" stands for "Yes, the patient has the disease", "N" stands for "No, the patient does not have the disease", "Q" stands for "it is questionable whether the patient has the disease", and "U" stands for "disease is not mentioned in the record". "Y", "N", "Q", and "U" are all valid judgments for textual annotations. But only "Y", "N", and "Q" are valid judgments for intuitive annotations.

Posted on 03-18-08

Q: Not all of the records have judgments for all of the co-morbidity/source combinations. Are these missing judgments supposed to be left out of the training data, or are they to be treated as "unmentioned" ("U")?
A: We have deliberately left out some of the records from the training set for some co-morbidities. This is necessitated by lack of agreement for those co-morbidities on those records. In other words, there is no judgment for the records that have been left out for a co-morbidity/source combination. Please exclude them from training.

Q: But there are so many missing records!?
A: Yes, especially for the intuitive judgments, the agreement was low. As a result, a lot of records are excluded from this task.

Q: But for testing purposes, how are we going to know which records to provide judgments for and which ones not to bother to classify for each co-morbidity?
A: You will be provided with a template that will list the record id's for which you are asked to provide judgments.

Q: What is the relationship between records marked "U" and records that are missing?
A: Records are labeled as "U" (Unmentioned) only if the annotators agreed that the co-morbidity was not mentioned in the patient record. This is distinctly different from cases which are omitted due to disagreement, e.g, one annotator marked the record as "U" and the other marked it as "Y" (and the third annotator failed to break the tie.) The missing documents represent the disagreements.

Posted on 02-07-08

Q: Are the data longitudinal (same patients over time), or are they better characterized as a snapshot of different patients' records?
A: The records are longitudinal to the extent that each record provides longitudinal information. We are using one record per patient.
Correction: In the data, some patients are represented by multiple records. But the identifiers have been removed such that record linking is not supported.

Q: What is the goal of the obesity challenge?
A: The high-level goal is to encourage the development of systems that facilitate studies of obesity by grouping patients based on obesity and co-morbidities they exhibit. We hope we will then be able to use these systems to gather data and study correlations on diseases, for example.

Q: How is the obesity challenge different from last year's smoking challenge?
A: On the technical side, this is a follow-up to the smoking challenge. One difference is that smoking challenge included a time component to classification. In this case, negations and uncertainty are to be explored. Also, this year, we will be able to release intuitive judgments which are based on and supported by the text but may require interpretation that goes beyond explicit assertions.

Posted on 02-05-08

Q: Can the challenge participants use external data sources (in addition to the challenge training data) in developing their systems?
A: We will allow use of external sources but will keep the evaluation of systems that use external sources separate from those that do not.

Q: In the training data, are the annotations of obesity and co-morbidities inline xml, or standoff? The web page indicates that the submission format is standoff, but does not characterize the training data format.
A: Everything is standoff.

Q: Are the annotations assigned at the document level, or are they applied to text elements that support such classifications?
A: Annotations are at document level. The evidence to support the classification has not been marked.

Q: Are the data longitudinal (same patients over time), or are they better characterized as a snapshot of different patients' records?
A: The records are longitudinal to the extent that each record provides longitudinal information. We are using one record per patient.

[back to top]