Cold Start 2012
Frequently[1] Asked Questions

Version 1.2 of June 29, 2012

Question: I'm very concerned about the GPE slots. Our system doesn't have any functionality to generate these slots. Nor is there training data.

Answer: If you are worried about slots of PERs or ORGs that have GPE fillers, then there is training data for the those slots from 2010 and 2011 that are string-filled. There is also plenty of training data for entity linking of GPEs. There is no provided training data for the two together, but the ability to train the components separately should suffice.

If on the other hand you mean slots that are properties of GPEs (e.g., `gpe:births_in_country`), there are no such slots in Cold Start that are not the inverse of a slot that is already in the 2011 and 2012 slot filling task. So an existing slot filling system will suffice for the task.

Question: The confidence measure section is still vague. Can you provide more extensive examples so that participants have a better idea about how to generate comparable confidence values?

Answer: There is no requirement that confidence scores be comparable this year, nor even a requirement that they be present. We don't currently know enough about how confidences should be expressed and used to provide suggestions. Rather, we hope that one or more groups will do something sensible with confidence values, and we will learn from that how to measure performance with confidences in future years. All we claim for this year is that if confidences are present, we will use them in two ways. First, we will use them to induce an ordering on the assertions, then provide some additional measures on the ordered list of assertions. For example, instead of simply giving a single recall/precision point, we might give the full recall/precision curve. Second, if the same triple is submitted more than once, or if a single-valued slot is given multiple fills, we will use only the candidate with the highest confidence value. We welcome feedback on the best ways to exploit confidence measures.

Question: Will KB evaluation queries appear in natural language that requires a system to do automatic parsing? Or this will not be part of this year's requirement?

Answer: No, there is no such requirement. In fact, neither the participants nor their systems will see the evaluation queries in any form prior to the evaluation. Rather, each system will produce a knowledge base that is then fed by NIST to a knowledge base engine such as Stardog; it is the KB engine that will be given the evaluation queries. All evaluation queries will be expressed as SPARQL queries when they are fed to the KB engine (although we will likely generate them using a much simpler format).

[1] Defined as 'at least once, plus or minus one time.'

Question: How do we indicate that a given slot has a NIL value?

Answer: There is no explicit representation for NIL slots in Cold Start 2012. Simply do not assert any value for a slot that should be NIL.

Question: What happened to entry points?

We have changed the task specification with regard to entry points. A mention string for an entity might be used as an entry point (i.e., as evaluation input), or it might be used as a slot fill (i.e. as evaluation output). We could put together a list of entry points, but we cannot guarantee that all mentions in that list are correct (as to tag the entire collection for named entities must be automated). Yet participants are required to ensure that the strings they use for slot fills are correct. Thus, they must be given the freedom to modify any mention in the entry points list. Since they may in doing so modify an actual entry point, providing an official entry points list doesn't guarantee us that all entry points will be defined in each submission. The upshot is that we will distribute a list of named entities, but that it will be a suggestion rather than a requirement. Instead, a single character in a document will be used as the entry point for an evaluation query. Whatever submitted mention includes that character will be used to identify the starting node in the submitted KB. Because mentions are now required to be non-overlapping, this approach will identify at most one starting point for each evaluation query.

Question: Why are pronouns not considered mentions? For example: Entity039 – “Maggie Simpson” in SIMPSONS016 - appears in the same document as “she" (- starts to cry).

Answer: They are mentions, but they are not named mentions. For this year at least, TAC will use only named mentions. You are welcome to identify and use nominals and pronominals in your processing (indeed doing so will probably be necessary to score well), but they will not appear as mentions in the KB.

Question: From what character exactly should the offset in the provenance be counted? It seems that it starts from the start of the “<DOC>” tag, which is counter-intuitive (we would expect it to start from the beginning of the text, so that we can use a standard XML reader to extract the text, and ignore all XML tags around it).

Answer: LDC collections often contain important information that is not part of the text (headlines, datelines, etc.). So we have to allow offsets prior to the <TEXT> tag. Whether the tags themselves are counted in the offsets is a tougher question. The ACE evaluation did not include the tags, and it performed various other transformations such as mapping <CRLF> to a single character. Some felt those rules complicated processing, but others preferred them. Nonetheless, all of the LDC tools calculate offsets on Unicode characters, without any preprocessing. Because LDC is providing the document collections, the evaluation queries and the assessment, it is important to maintain compatibility with their tools.