AQUAINT Pilot Knowledge-Based Evaluation: Annotation Guidelines

Dick Crouch, Roser Saurí, Abraham Fowler, May 2005

Introduction

This document sets out some guidelines for annotating textual inferences for the AQUAINT knowledge-based evaluation pilot. The annotated data can be used in a variety of ways to support a number of different evaluation scenarios. Not everything in the annotation will directly reflect something that is either a system input or output. Some of the annotation fields are present solely to allow different ways of decomposing and analyzing system results. While it is not the purpose of a set of annotation guidelines to specify a particular system or evaluation scenario, we feel that it may nonetheless help to first sketch out one possible vanilla system scenario.

Possible System Inputs and Outputs

A system is presented with a passage of text and a question about the passage. The system gives a response to the question, and indicates the "force" of the response. The force can either be *strict* (assuming the truth of the text, the response inescapably follows), or it can be *plausible* (assuming the truth of the text the response is reasonable, although additional information might have led to a change in the response). In addition, the system can offer (a) a system-internal justification for the response (which data-sources, forms of reasoning, etc were used to produce the response), (b) a human readable explanation of why the response is a correct answer to the question, and (c) an indication of the system's confidence in its judgment of the response and its force.

- System output:
  - Mandatory:
    - Response
    - Force: Strict / plausible
  - Optional
    - System justifications (e.g. linguistic/world-knowledge)
    - Human readable explanations
    - System confidence
    - … other possible outputs (e.g. contexts)

Annotation Fields

A richer set of annotations is given on development / evaluation material to allow for more informative ways of breaking down results. In addition to the text passage, the question, the response, and the force of the response, annotations must also state the "source" and "polarity" of the response. The source (*linguistic* or *world-knowledge*)

indicates whether or not the question can only be answered by reference to additional world knowledge that is not made explicit in the passage and question. The polarity (*true*, *false*, or *unknown*) indicates whether the response is true, or false or unknown (assuming the truth of the passage) Further optional annotations may (a) give a textual characterization of any additional world knowledge needed to derive the response, (b) specify other assumptions made in deriving the response, especially whether a particular interpretation of the passage or question was assumed in cases where ambiguity affects the outcome, (c) whether the response invokes a particular context, such as what is believed, or what is planned, (d) the annotator's confidence in their judgments, and (e) the provenance of the passage and/or question, e.g. from actually occurring text, simplified from actual examples, or hand-created example.

- Annotation:
  - Mandatory:
    - Passage
    - Question
    - Response
    - Polarity: True / False / Unknown
    - Source: linguistic / world-knowledge
    - Force: strict / plausible
  - Optional
    - Characterize additional knowledge that the response depends on
    - Assumptions (including which interpretation if ambiguous)
    - Context-type (belief, plan, …)
    - Annotator's confidence
    - Provenance

To reiterate, the annotations contain more information than the QA system is likely to produce. This is for two reasons. (a) To allow the same annotated date to be used in evaluating a variety of different systems with different levels of output. (b) To allow various forms of error analysis on a single set of evaluation results. For example, when a system produces an incorrect response, how often does it go wrong because of incomplete world knowledge, and how often because of basic misunderstanding of the passgage or question? How often does it produce answers that are demonstrably false, as opposed to answers that are merely not justified by the passage? Does it do particularly well or badly when looking at particular context types?

Examples

The following examples illustrating the general meanings of the polarity, force and source annotations are briefly discussed below.

```
1. PASSAGE: Queen Victoria died in 1901, at Osborne House on the Isle
   of Wight.
   QUESTION: Who died in 1901?
   RESPONSE: Queen Victoria
   POLARITY: True
   FORCE: Strict
   SOURCE: Linguistic
```

2. PASSAGE: Queen Victoria died in 1901, at Osborne House on the Isle of Wight.
   QUESTION: Who was born in 1901?
   RESPONSE: Queen Victoria
   POLARITY: False
   FORCE: Plausible
   SOURCE: World
   BECAUSE: Most people, especially crowned monarchs, do not die in the year they were born

3. PASSAGE: Queen Victoria died in 1901, at Osborne House on the Isle of Wight.
   QUESTION: Who was born in 1901?
   RESPONSE: Queen Victoria
   POLARITY: False
   FORCE: Strict
   SOURCE: World
   BECAUSE: Queen Victoria died at the age of 86.

4. PASSAGE: Queen Victoria died in 1901, at Osborne House on the Isle of Wight.
   QUESTION: Who was born in 1901?
   RESPONSE: Queen Victoria
   POLARITY: Unknown
   FORCE: Strict
   SOURCE: Linguistic

5. PASSAGE: Queen Victoria died in 1901, at Osborne House on the Isle of Wight.
   QUESTION: Who was born in 1901?
   RESPONSE: Don’t know
   POLARITY: True
   FORCE: Strict
   SOURCE: Linguistic

6. PASSAGE: Queen Victoria died in 1901, at Osborne House on the Isle of Wight.
   QUESTION: Who died in 1901?
   RESPONSE: Prince Albert
   POLARITY: Unknown
   FORCE: Strict
   SOURCE: Linguistic

7. PASSAGE: Queen Victoria died in 1901, at Osborne House on the Isle of Wight.
   QUESTION: Was Queen Victoria born in 1901?
   RESPONSE: No.
   POLARITY: True
   FORCE: Plausible
   SOURCE: World
   BECAUSE: Most people, especially crowned monarchs, do not die in the year they were born

Example (1) is a case of a strict entailment, where the response to the question follows directly from the passage without any need for additional world knowledge.

Example (2) is a case where the response is false given the passage, although it requires some additional world knowledge and plausible reasoning to detect this. The additional knowledge might for instance be that few people, and especially few crowned kings and queens, are born in the same year as they die.

Example (3) illustrates how using a different set of world knowledge can change the force of an answer. Knowing that Victoria lived to be 86 allows one to conclude strictly that she was not born in 1901, the year she died. Although it is officially optional to characterize additional knowledge used, this annotation is very strongly recommended when the source is world knowledge.

Examples (4) and (5) consider the same question and response as (2) and (3), but from a stand-point where no world knowledge is assumed. Example (4) is an instance of a distractor, where a response that a QA system should not get is explicitly pulled out: the response is a wild guess that is neither verified nor falsified by the source passage Example (5) illustrates the kind response that one would hope for from a QA system: rather than make a unverifiable and unfalsifiable wild guess, the correct response of "don't know" is given. The annotation scheme is deliberately set up to allow for two different kinds of distractor: ones like example (3), where the response is demonstrably false, and ones like example (4), where the response is merely unjustified. These distinctions are very hard to make (though not, strictly, impossible) given evaluation material that only records completely correct responses.

Note also, in going from examples (2) and (3) to examples (4) and (5), the effect of factoring out world knowledge. When a true/false response is dependent on world knowledge, the effect of removing the additional knowledge will be to change it to an unknown response.

Example (6) is a slightly contrived case where the passage does not provide enough information to determine whether the response is true or false, and so the response "Prince Albert" is marked "unknown" on the basis of linguistic knowledge. However, given enough additional world knowledge, of a historically very specific sort, the response could also be false: Queen Victoria famously mourned the death of her husband Prince Albert for many years, and so if she died in 1901, he must have died well before then. But this degree of specificity is inappropriate, and responses should be based on general, common-sense world knowledge (see below). Example (6) is somewhat contrived, since without this specific world knowledge there would be no reason to link Prince Albert's death to that of Queen Victoria's. As a general rule, unknown responses should be confined to entities or facts that are explicitly mentioned in the passage (the same does not hold for true or false responses). Example (6) violates this general rule. While the annotation is correct, it is an example of a passage/question/response pairing that should be avoided.

Example (7) is the analog of example (2), where a wh-question has been replaced by a yes-no question.

Polarity: {True, False, Unknown}

Definitional question: *What does the passage imply about the truth or falsity of the response to the question?*

Assume that the passage is true (suspending disbelief if you happen to know that it is not). If, on this assumption, the response is correct, mark its polarity as **true**. If the response seems not only incorrect, but in fact contradicted by the passage, mark the response's polarity as **false**. Otherwise, if the response seems incorrect because you cannot tell whether it is implied or contradicted by the passage, mark its polarity as **unknown**. In all these three cases, the Source of knowledge is of Linguistic type (see the section on Source below).

Alternatively, for cases of unknown polarity like those described above, you can also use your knowledge about how things are in the world and mark the polarity of the answer as **true** or **false** (whatever you know it is the case), and indicate World Knowledge as the value of the Source attribute.

Guidelines
As just mentioned and illustrated in examples (4) -- (6) above, additional world knowledge can change a linguistic-based polarity of unknown to either true or false. It can also change world knowledge-based plausible polarities to any of strict false or true. Where additional world knowledge can change a polarity, it is perhaps advisable to duplicate the example, giving one annotation with and one annotation without the additional knowledge.

```
PASSAGE: Queen Victoria died in 1901, at Osborne House on the Isle
of Wight.
QUESTION: Who was born in 1901?
RESPONSE: Queen Victoria
POLARITY: Unknown
FORCE: Strict
SOURCE: Linguistic

POLARITY: False
FORCE: Strict
SOURCE: World
BECAUSE: Queen Victoria died at the age of 86.
```

Another difficulty will be in distinguishing whether the truth/falsity of the conclusion really is a consequence of the premises, and not just something that is known anyway. For example, if one happens to know that Queen Victoria was born in 1814, it is tempting to mark the conclusion "Queen Victoria was born in 1901" as false, whatever the premise text. One way of determining whether the conclusion is a genuine consequence of the premises is to consider variations on the premise text, and ensure that at least some of the

variations eliminate or weaken the conclusion or change its polarity. If the conclusion never changes, then it is not a consequence of the premises. Consider

```
PASSAGE: It is irrelevant that there was terrorist activity in
Baghdad.
QUESTION: Was there terrorist activity in Baghdad?
RESPONSE: Yes
POLARITY: True
SOURCE: Linguistic
```

How can you disentangle your prior knowledge that there has been terrorism in Baghdad from the positive response to this particular textual question? Try changing the word "irrelevant" to "unlikely." If your response and polarity remain the same, then perhaps your judgments are not being driven by the contents of the passage.

An understandable confusion about the polarity field is as follows: why is it necessary to give both the response, and the polarity of the response? Especially for yes-no questions, isn't the polarity (true/false/unknown) completely redundant given the response (yes/no/don't know)? And for wh-questions, what exactly is the purpose of recording incorrect answers? The purpose is to allow for the possibility of more detailed error analysis on the results of evaluation by including distractors in the evaluation material. There are two ways of getting an answer wrong: (i) giving an answer that is just unsupported given the data, and worse (ii) giving an answer that is just plain false given the data. If the evaluation material only contains examples of correct answers, it will be hard to distinguish these two forms of failure. For training purposes, one would probably want to ignore distractors (polarity = false or unknown). But at present, we are only engaged in producing evaluation and development material, not large quantities of training material.

Force: {Strict, Plausible}

Definitional Question: *For responses with polarity true or false, could additional information (consistent with the passage) make you change your mind about the polarity?*

A strict inference is one where in all circumstances in which the premises are true, the conclusion has to be true (alternatively, false). A plausible inference is one where it is reasonable to conclude that the conclusion is true (alternatively, false), even though under certain special circumstances in which the premises are true, the conclusion might not be true (alternatively).

Guidelines
Strictly speaking, a response labeled as plausible only makes sense with true or false polarities. Cases with the polarity evaluated as unknown indicate that there is not enough information to answer. Those cases should not be confused with examples with true or false polarity and plausible force, since the polarity value of these can effectively be

decided on the basis of partial information, although additional information may cause to change it.

To judge if a true or false response is a plausible inference, add the contrary of the response to the passage, and see if the passage can still coherently be taken to be true. For example:

```
PASSAGE: Queen Victoria died in 1901, at Osborne House on the
Isle of Wight.
QUESTION: Who was born in 1901?
RESPONSE: Queen Victoria
POLARITY: False
FORCE: Plausible
SOURCE: World
```

Consider the augmented passage: "*Queen Victoria died in 1901 at Osborne House on the Isle of Wight. She was also born in 1901*." This is a coherent (albeit false) passage, indicating that a plausible inference was used to derive the response.

By contrast, consider

```
PASSAGE: A rocket attack killed two Israelis, including a 3-year-
old boy.
QUESTION: Was a three year old killed in a rocket attack?
RESPONSE: Yes
POLARITY: True
FORCE: Strict
SOURCE: Linguistic
```

The augmented passage "*A rocket attack killed two Israelis, including a 3-year-old boy. But no three year olds were killed in a rocket attack*." is incoherent, indicating that the inference is strict.

Some care needs to be taken in forming the contrary of a response. Simply adding (or removing) the word "not" will not, in general, suffice. "A three year old was not killed in a rocket attack" does not necessarily contradict "A three year old was killed in a rocket attack", since there can be a consistent reading where one three year old was killed, and another three year old wasn't.

When the force of a response (either strict or plausible) depends on world knowledge, it is important for the annotator to attempt to encapsulate this in the field characterizing additional knowledge. This encapsulation should be in textual form, such that if it were added to the source passage, then ideally no additional world knowledge would be required to judge the force of the response. This ideal, of stating all the required additional knowledge, will often not be achievable. So annotators should aim at providing enough additional text to justify or explain their force judgment to other annotators or human readers. Or put another way, the additional knowledge should be specified in the same kind of textual form as the source passage, so that when they are taken together a person of normal intelligence but perhaps unusual ignorance should agree on the judgments of force and polarity.

Source: {Linguistic, World-knowledge}

"Definitional" Question: *Would any competent speaker of the language, no matter how ignorant otherwise, make the same judgment about the polarity of the response?*

A linguistic/lexical inference depends solely on the construction of and on the meanings of the words used in the premise and conclusion texts. An inference is based on world knowledge if it additionally requires some knowledge of the world, either basic facts (e.g. France is in Europe), or more extensive chains of reasoning about the world.

Perhaps the hardest thing to determine is whether world knowledge is required in inferring the conclusion; the dividing line between linguistic/lexical and world knowledge is notoriously hard to draw. This is why "definitional" is in scare quotes above. Some tests for detecting dependence on world knowledge are as follows:

1. If the conclusion follows directly from the premises without appeal to further intermediate premises or lemmas, then the inference is linguistic/lexical.

2. If the conclusion does require intermediate lemmas, and these refer to spatially or temporally specific objects that are not mentioned in the premise text, then the inference relies on world knowledge. For example, that "Smith visited Baghdad" implies "Smith has been in Iraq" relies on the spatially specific lemma "Baghdad is in Iraq", and Iraq is not mentioned in the premise text.

3. If the premise and conclusion texts allow (all possible) uniform replacements of objects of the same 'type', then the inference is linguistic/lexical. For example, replacing "Queen Victoria" by a term referring to some other person, e.g. "the very premature baby", would break the inference, indicating that it cannot be linguistic/lexical. This test presupposes some useful typology of objects (people, places, dates, animals, artifacts, feelings, etc.) It is possible that the categories used in the WordNet lexicographer files could form the basis of a rough and ready typology.

A finer grained classification of source is possible, e.g :

LEX	Lexical information such as lexical chains
SR	Semantic relations
LNG	Linguistic knowledge
BWK	Basic world knowledge
XWK	Extended world knowledge

Where LEX, SR and LNG partition the linguistic/lexical class, and BWK and XWK partition the world knowledge class. But this finer grained classification is left as optional for now, and we will not attempt explicit guidelines.

Because: Characterizing additional knowledge

Additional knowledge required to justify a response should be encoded textually, for example:

```
PASSAGE: Congressman Smith visited Baghdad.
QUESTION: Has Smith been to Iraq?
RESPONSE: Yes
POLARITY: True
FORCE: Strict
SOURCE: World knowledge
BECAUSE: Baghdad is in Iraq.
```

Additional annotation along these lines is problematic in cases were relatively complex chains of inference about world knowledge are required. This is because there can often be several alternative chains, using different items of world knowledge. Textual paraphrase explicating purely linguistic chains of inference are likely to be hard to produce. But for inferences based on world knowledge, you should aim to characterize the additional knowledge / facts / premises assumed, though without attempting to spell out the chain of inference that links these facts together to establish the conclusion.

Ideally, the additional knowledge should be written down in such a way that, if its description were append to the text passage, then the response would follow without need of any further unspecified world knowledge. As mentioned previously, this ideal will not always be achievable. One way of thinking about the text to go in this field is that it should just state the facts relevant to the response that the source passage happened to leave out. Another way of thinking of it is as a way of reporting to other annotators the facts relevant to your own judgment of force and polarity.

This field should *not* be regarded as an open-ended, free-form comment. Such comments properly belong as part of the assumptions field, described below. Rather, it should be viewed as an extension of the source passage pointing out the perhaps obvious, so that even someone who is very ignorant could understand why the answer was right or wrong, plausible or strict. It is possible, though by no means a foregone conclusion, that the annotations in this field could in the future form a partial basis for evaluating system produced, human-readable answer justifications.

Although this annotation field is (at present) only optional, we strongly recommend that annotators make every effort to fill it in cases where the source of a response is world knowledge. This is because a single passage/question/response can receive multiple, correct, judgments of force and polarity, depending on the amount and nature of additional world knowledge assumed (cf. examples (2) and (3). This is particularly so in the case of plausible responses, where further facts can render a once plausible response strictly false (or strictly true). It is important that the dependency of force and polarity on additional knowledge is explicitly recorded.

Further Assumptions

Some responses depend on a particular interpretation of the text passage or question. When more than one possible interpretation is detected, any assumptions that are tied to a particular interpretation and lead to a particular Response, should be annotated in this field along with the Response. The assumptions can be written in an unstructured, plain English form. In some cases it may be possible to describe which reading of the passage or question is assumed, but often this degree of precision will be difficult. In lieu of this, when ambiguity affects judgments, this can be indicated as follows:

> `One reading`: The judgments about the response hold under at least one interpretation of the passage and the question
> `Preferred reading`: The judgments about the response hold under the most likely interpretations of the passage and the question
> `All readings`: The passage and/or question are ambiguous in ways that could have a bearing on the response, but in fact the judgments about the response hold under all interpretations that the annotator can see.

However, if you feel that you are able to describe and pick out the particular interpretation under which the annotation judgments hold, then by all means describe it in this field, as a free-form comment. Unlike the additional knowledge field above, there is no reason for the comment not to contain such linguistic terms of art as "on the reading where the postverbal PP attaches low," and the like.

It is probable that this field will turn into an area where decisions resolving (resolvable) inter-annotator disagreements are recorded. Disputes can and will arise because different annotators work on different interpretations of the passage and question. For example

```
PASSAGE: Gil welcomed the release of prisoners
imprisoned during the October riots.
QUESTION: Have prisoners imprisoned during the October
riots been released?
RESPONSE: Yes
ANNOTATOR1: polarity=true, force=strict,
source=linguistic
ANNOTATOR2: polarity=unknown, force=strict,
source=linguistic
```

The probable cause for this disagreement is an ambiguity in the passage, depending on whether Gil is welcoming a planned future release of prisoners (annotator2), or is welcoming a release that has already happened (annotator1). Oftentimes, it will not become apparent that there is an ambiguity until such disagreements occur.

How can one tell that a disagreement arises through ambiguity? It is not possible to give hard and fast rules. But one characteristic is that the disagreement will persist despite trying to bring in more and more additional world knowledge to sway the decision one way or the other. For a non-ambiguous example, like the Queen Victoria died in 1901

passage, one can imagine an annotator starting out by judging it as false, on strict linguistic grounds, that she was born in 1901. Others might try to change the annotator's mind by imagining alternative scenarios, e.g. suppose that Victoria had been an infant queen who died young. If these scenarios succeed (as in this case they should) in swaying the annotator, then ambiguity was not the source of the disagreement. If on the other hand, the annotator remains invariant through alternative scenarios, or rejects many of them as being plain incompatible with the source passage, then chances are there's an ambiguity lurking somewhere. Unfortunately, it sometimes requires quite a degree of linguistic skill to pinpoint the source of the ambiguity.

In summary, when annotators do not agree on Response, Polarity or Force, one first step is to have each one explain exactly how they understood the Passage and Question. It helps to break long Passages and Questions into shorter, simpler sentences with fewer structural attachments that can be interpreted in multiple ways. If the passage and question are interpreted the same way by the annotators, it is likely the difference lies in the *knowledge* used to arrive at their differing annotations. In this case, both annotators' Response/Polarity/Force can be included in the final corpus, each justified by the different world knowledge used in arriving at them. The two annotations will differ in their Because fields. If, however, there is disagreement on the meaning of the Passage and/or Question, then the source of disagreement likely lies there. In this second case, both annotators' Response/Polarity/Force can still be included in the final corpus; their Because fields may still differ; but in addition, the Assumptions fields will need to provide enough information to lead to the exact interpretation each annotator chose. In other words, the Because field will contain knowledge used in reasoning to arrive at the Response/Polarity/Force, while the Assumptions field will contain information used to *disambiguate* the meaning of the Passage or Question.

Context Type

When a response is not marked as strictly true, there can be specific contexts in which the response does come out as strictly true. For example

```
PASSAGE: The book claimed that Queen Victoria died in
1901, at Osborne House on the Isle of Wight.
QUESTION: Who died in 1901?
RESPONSE: Queen Victoria
POLARITY: Unknown
FORCE: Strict
SOURCE: Linguistic
CONTEXT: Report, polarity=true, force=strict,
source=linguistic
```

That a book claims something does not necessarily make it true. But in the context of what is reported by the book, the claim is presumed true. Other contextual-reasoning modules will be able to determine the default credibility of certain contexts, and thus determine the circumstances under which a contextually limited response can be elevated

out of its context. For example, that the book in question is a reliable source of information, so that anything it claims to be true should be taken as true. Sometimes there may be multiple contexts in which the response is true.

When a report context is identified, the annotation of polarity, force and source (at least) should be repeated, giving the values that apply within the context. Sometimes several contexts can be invoked by a passage or question, and in these cases multiple context annotations can be given.

It is too early to impose a complete list of different context types, as the length of the following discussion indicates, but here is a preliminary list:

- REPORT: contexts in which there is an agent reporting or informing about something that: (a) has supposedly happened or will happen, or has been decided (e.g., clauses introduced by predicates like 'recount', 'report', 'inform', report', announce', 'disclose', 'notify', 'tell', 'say', etc.), or (b) adds some information to a previous report (e.g., clauses introduced by 'add', 'answer', 'reply', 'respond'). Report contexts need not necessarily imply anything about what the agent believes or disbelieves, since agents can knowingly give false reports. However, reports do commit the agent to publicly maintaining the truth of the report, on pain of revealing themselves to be insincere. Sub-classes of report contexts include (Bach and Harnish, 1979):
    - Assertives: affirm, allege, assert, claim, maintain, state.
    - Confirmatives: conclude, confirm, diagnose, find.
    - Retractives: deny, disavow, disclaim, retract.
    - Dissentives and disputatives: differ, disagree, dissent, reject, dispute
    - Predictives: forecast, predict.
    - Suggestives: guess, hypothesize, speculate, suggest, conjecture.
    - Concessives: acknowledge, admit, concede, concur, confess
    - Assentives: accept, agree, assent

  Some report contexts commit the both the reporting agent and the author of the text to the proposition (e.g. concessives and assentives).
- BELIEF: Contexts that imply what the agent believes, for example clauses introduced by ‘believe’, ‘know’, ‘expect’. As with reports, some belief contexts also commit the author of the text to the proposition. To say that “Smith knows Hussein is in Paris” says both something about what Smith believes, but also commits me, as the author, to a claim about where Hussein is. Note that first person belief contexts such as “I know that Hussein is in Paris” can easily be confused with report contexts, since I (as the text author) can be insincere in reporting what I (as the agent) know.
- VOLITIONAL: contexts expressing the hopes, fears or intentions of the agent. Most of these contexts, especially when the proposition has future time reference, do not commit either author or agent to the truth of the proposition (hopes, wishes, intends).
- PLANNING: contexts expressing what is planned or scheduled (“The train arrives at 4pm”, “They agreed to meet next Tuesday”).

- COMMISIVE/DIRECTIVE: contexts involving giving permission or orders, including
    - Requestives: beg, ask, petition, plead
    - Requirements: command, demand, instruct, order, require
    - Prohibitives: forbid, prohibit.
    - Permissives: allow, authorize, consent, grant, license.
    - Advisories: advise, warn, recommend.
    - Commisive: propose, offer
- POSSIBILITY: contexts of probability or likelihood ("It may rain tomorrow", "He is unlikely to know Hussein").
- CONDITIONAL: contexts conditional on some antecedent condition ("If Jim wins the next round, he wins $100,000", "Prompt action prevented a disaster; if the guard had not been alert there would have been fatalities.")
- TEMPORAL: contexts limited to specific times (Before dinner, Jim is cranky.)
- SPATIAL: a spatial context is defined (In Texas, summers are very hot.)
- DOMAIN: a context restriction on the domain. This can be for unstated information as well as word/concept senses. (Reggy Miller shot the lights out. Domain: NBA vs. Domain: Law Suits)

Annotator's Confidence

The annotator's confidence in their own judgments should not be confused with their judgments about the polarity or force of the response. That is, a plausibly true response should be annotated as such, and not as a strictly true response with only a moderate level of confidence. A three point scale (absolutely confident, fairly confident, not confident) is probably sufficient.

Provenance

Evaluation material can contain a mix of actually occurring and hand-constructed passages and questions. Hand constructed examples are permitted as controlled, "laboratory" specimens to allow evaluation or debugging/development to probe into system performance in a particular area (e.g. how well does it do on temporal inferences, presuppositions, belief contexts, etc). Actally occurring examples are of course important for indicating how well the system might do in the field. To allow natural and hand-crafted examples to be distinguished in evaluation, it is useful to record the provenance of the examples: either (i) **hand-crafted** for completely made up examples, (ii) **edited(*source*)** for examples edited from an actual source, and (iii) **unedited(*source*)** for unedited examples from an actual source. In (ii) and (ii), *source* should specify the source of the text. Since most questions and responses will probably be hand-crafted, it is only necessary to record the provenance of the text passage.

Non-Orthogonalities

As far as possible, different annotation/evaluation dimensions should be independent and orthogonal. In this section, we indicate some of the places where there are interactions between the annotation dimensions.

Polarity and Force: A plausible response is one where you jump to a conclusion on the basis of partial information, but don't rule out the possibility of revising your conclusion when more information becomes available. An response of unknown means that you don't have enough information to decide one way or the other, and are not willing to jump to a conclusion. In other words, you shouldn't mark something as having jumped to a plausible conclusion of "unknown". More specifically, if force=plausible, then you should not have either of the annotations (a) polarity=unknown, or (b) response=don't know, & polarity=true.

Source and Polarity: It is hard to see the motivation for annotating a response that remains unknown even when world knowledge has been added. Thus, the dimensions source=world & polarity=unknown, while not inconsistent, are of doubtful utility.

Selection of Questions and Responses

While not strictly part of the annotation, compiling evaluation material will typically involve hand construction of the question and response, if not also the passage. A few comments about the selection. First, the question should be answerable on the basis of the passage, and the information in the passage should be necessary/required for answering the question. When selecting responses that are annotated as unknown, the response should mention entities and propositions that are at least mentioned in the passage and/or question (thus ruling out the Prince Albert example (6)). In general, one would expect the evaluation material to focus on responses that are true (either strictly or plausibly) given the passage, with false and unknown responses being present to throw in some distractors. We leave the appropriate ratio of distractor to correct resonses open. When a question has several alternative correct responses (usually, wh-questions), we also leave it open as to whether or not an exhaustive responses should be produced.