Why Toronto?

The Final Jeopardy question for game one of the two-game tournament between Watson, Jennings, and Rutter, in the category, "*U.S. CITIES",* was*, "Its largest airport is named for a World War II hero; its second largest, for a World War II battle."*

Watson gave "Toronto" as its answer. The correct answer, "Chicago", was a very close second, however both answers had low confidence (around 30%). During regular game play, this would have been well below the threshold to "buzz"; Watson is effectively saying, "I don't know", but for Final Jeopardy! questions you must give an answer and if there is a 30% chance of getting it right, you go with that.

The immediate response from people, is, "Why does Watson think Toronto is a U.S. City"? Really, a better question is, "Why does Watson think Toronto is a better answer than Chicago, even if it is unsure?" Watson considers a lot of answers and evidence, it is not always the case that the type, which may be clear to people but often unclear or ambiguous to Watson, is the deciding factor.

In the Jeopardy! setting you have no opportunity to explain yourself, but as an analytic tool used in some practical task, like diagnosis, the question-answering technology that underlies Watson does provide explanations in the form of *evidence profiles*. An evidence profile gathers together the evidence for an answer into different dimensions, such as type, geospatial location, time, popularity, string match, passage support, etc. It is often through exploring evidence profiles for *wrong* answers that you can better understand what Watson is doing.

A closer look at the evidence profile for the answer "Toronto" in this clue reveals a few surprising things. Not surprisingly, Toronto scored relatively well in the string match dimension: string match is a collection of evidence based on more traditional types of search, that do not consider what questions mean, but simply try to match the words in the question to words in a document. There are articles about Toronto that mention "largest airport", "named for a World War II hero" (not the airport, but places in Toronto; string match does not consider what it means, just that the strings are there), "second largest", and "battle". In fact, in this dimension, Toronto has stronger evidence than Chicago, because it has been more frequently associated with these strings in Watson's sources.

The fact that this kind of evidence can be unreliable is precisely why Watson must go beyond web-style search to get answers, and most of the time this other evidence helps. One of the evidence dimensions that people immediately sense is the type dimension: Toronto is not a "U.S. City". At first, team members suspected that Watson had merged evidence from different cities named Toronto – this sometimes happens as it is often difficult to pick apart different things with the same name, and there are eight cities in the U.S. name Toronto. This may have happened to some degree, but was not the main cause of the failure. Let's look deeper.

The type dimension of evidence begins by taking the unmodified type word from the clue, and this is "cities". So the primary form of type evidence does not consider the "U.S.", just the "city", and clearly Toronto scores highly as a city, as does Chicago.

Most people probably don't even notice this, but syntactically the "U.S." in this clue is a *modifier* to the type word "cities", and while Watson does consider this type modifier, often in Jeopardy! clues type modifiers are not that useful, for example "barbarian city" (according to whom), "southern city" (south of what), "colonial city", etc. In fact, the very cause of the failure here is an example of why type modifiers are a weak source of evidence: Watson is looking to see if any of its sources use the same modifier, and in fact one source calls Toronto an "American League City", because the Toronto Blue Jays play there, and since Watson believes American to be synonymous with U.S., it believes it has found evidence that Toronto is a U.S. City.

So, surprisingly, Watson thinks Toronto is a U.S. City. It is important to note that for this dimension of evidence gathering (matching modifiers), Watson does not really consider what U.S. and American *mean* – e.g. that they are countries that contain cities – rather it is treating them as synonymous words independent of meaning. You can imagine if someone had asked you to find a "brzznap city," and you found something that said, "finplap is another word for brzznap", and somewhere else you found "Toronto is a finplap league city", you would probably conclude Toronto is a brzznap city, even though you don't know what brzznap means. Not surprisingly, Watson also finds evidence in this dimension for Chicago.

A further complication for Watson in this case is that the type and modifier come from the Jeopardy! category, not directly from the clue. In fact, the clue does not explicitly indicate the type. Category names do not always indicate the type of the answer, rather, when they carry any meaning at all, they more commonly act as a topic heading. For example, a clue in the "U.S. Cities" category could have been, "Chicago's largest airport is named for this WWII Hero," in which the type would be "hero" with "WWII" as a modifier. This impacts all Watson's answers to the question by uniformly lowering their confidence, so in this case, since the top answer is incorrect, that is the right thing to do.

The fact that "U.S. Cities" is in the category and not the clue also impacts the part of the system that collects geospatial evidence. This evidence dimension actually considers what "U.S. City" *means* (as opposed to the type modifier evidence, which only considered them as words), and using geospatial databases is capable of concluding that Chicago, Illinois is in the U.S. and Toronto, Ontario is not. However since the "U.S. City" was in the category, not the clue, Watson wasn't able to reliably make sense of this as a geospatial constraint, and no evidence was acquired.

Another obstacle for Watson is that the question is complicated syntactically, and Watson failed to properly decompose the question into its parts. For multi-part questions, Watson should try to answer the two parts independently, and then find an answer that is common to both. For the first sub-question, "*Its largest airport is named for a World War II hero",* Watson gets reasonable answers like Toronto, New York, and Chicago, and as we've seen the type evidence supports all three. However the second part of the question leaves out the word "airport" and "named for", so the sub-question is, "Its second largest, for a World War II battle". Watson's decomposition wasn't able to make that a meaningful question, and none of the second sub-question answers resemble any of the answers to the first. The fact that Watson was unable to synthesize a result out of the two sub-questions contributes strongly to its lack of confidence in any of its final answers.

In the end, Watson was correct to place such low confidence in its answer – not because it was incorrect, but because it had insufficient evidence to justify even "Chicago" as an answer. With supporting evidence from types and modifiers, the slight advantage given to "Toronto" by string match evidence gave it a higher final rank.

As an interesting post-script, a lot of the errors Watson makes are, like this one, fairly obvious to people, and a consequence of its approach to consider many different options – often an order of magnitude more than a person would. By the same token, many of the things it gets correct are not obvious to people. This is clearly a good property of a decision support tool, which is what is at the core of Watson. Consider that in a post game discussion Ken Jennings told me he didn't really know the answer, but "considered both Chicago and New York, and I knew LaGuardia was not a WWII battlefield." (LaGuardia is NY's second largest airport). Now imagine the following scenario: a person is given the question, "This American city's largest airport is named for a veteran of WWI and WWII". Not really knowing the answer, but making a similar educated guess as Ken, the person comes up with "NY" and "Chicago". But being unsure if O'Hare or Kennedy served in WWI, the person asks Watson, and sees the answer "Toronto". The person then realizes that, in some sense of "American", it is not an unreasonable answer, and explores the evidence to find out more.