Why Co-Occurrence Information Alone Is Not Sufficient to Answer Subcognitive Questions

Robert M. French & Christophe Labiouse
Quantitative Psychology and Cognitive Science
Department of Psychology
University of Liege, Belgium
{rfrench, clabiouse}@ulg.ac.be

[Reference to cite: French, R. M. and Labiouse, C. (2001). Why co-occurrence information alone is not sufficient to answer subcognitive questions. *Journal of Theoretical and Experimental Artificial Intelligence, 13*(4), 419-429.]

Abstract

Turney (2001) claims that a simple program, PMI-IR, that searches the World Wide Web for co-occurrences of words in 350 million Web pages can be used to find human-like answers to the type of "subcognitive" questions French (1990) claimed would invariably unmask computers (that had not lived life as we humans had) in a Turing Test. In this paper, we show that there are serious problems with Turney's claim. We show by example that PMI-IR doesn't work for even simple subcognitive questions. We attribute PMI-IR's failure to its inability to understand the relational and contextual attributes of the words/concepts in the queries. And finally, we show that, even if PMI-IR were able to answer many subcognitive questions, a clever Interrogator in the Turing Test would still be able to unmask the computer.

Introduction

French (1990, 2000a) has argued that a computer that was "disembodied" —i.e., did not have a body like ours, did not experience the world as we humans did, etc. — could not pass a Turing Test (Turing, 1950; see French, 2000b for a review). Central to this claim was that the use of *subcognitive questions* would unfailingly unmask the machine. These are questions that tap into our human-specific verbal, physical and social interactions with the world with our human bodies, with our human visual and sensory apparatus, and with all the cultural trappings in which we find ourselves. Now, before administering the Turing Test, the Interrogator, a particularly smart and knowledgeable individual, would go out into the population and collect answers to a series of questions, like;

On a scale of 1 (awful) to 10 (excellent) please rate:

- How good the name *Flugly* is for the name of a glamorous Hollywood actress:
- How good the name *Flugly* is for the name of an accountant in a W.C. Fields' movie:
- Banana peels as musical instruments:
- Lawyers as slimeballs:

etc.

People's average ratings for the subcognitive questions would be used to produce a Subcognitive Profile that the Interrogator would use during the administration of the Turing Test. Whichever entity was farthest from the Profile would be the computer.

Ultimately, French (1990) claims that the Turing Test, rigorously administered, is too strong as a test for general intelligence. It turns out to be a culturally-specific test of *human* intelligence.

Turney (2001) claims that, while he agrees with French's position, he feels that the main point about a computer not being able to correctly answer subcognitive questions is wrong. He claims that "a simple unsupervised machine learning algorithm", PMI-IR that supposedly measures the semantic similarity between pairs of words or phrases can generate human-like answers to subcognitive questions. This program is based on the assumption that a word is "characterized by the company it keeps" (Firth, 1957) and measures the associative strength between two words by their average physical proximity over many millions of pages of text in the World Wide Web. Based on this measure of proximity, Turney then purports to show how PMI-IR can then give human-like answers to subcognitive questions, thereby demonstrating that "French is mistaken: a disembodied computer *can* answer subcognitive questions." It is unclear exactly why Turney would agree with French's position, given that the latter's position is based entirely on the ability of subcognitive questioning to unmask the computer.

In what follows we will make three points: First, we will show, empirically, that Turney's program PMI-IR simply does not work. We will further point out a number of shortcomings in the results of his program as presented in his paper. Second, we will give a number of simple theoretical reasons for the failure of PMI-IR. Finally, we will then show that, *even if his program worked*, it would still be possible for a clever interrogator to find subcognitive questions that would unmask the machine.

Failure of PMI-IR to develop a human -like Subcognitive Profile

Let us begin with a very simple example:

> "Rate on a scale of 1 (terrible) to 10 (excellent) the following: Lawyers as horses, fish, telephones, stones, sharks, cats, flies, birds, slimeballs, kangaroos, robins, dogs, and bastards."

We applied the PMI-IR search technique described in Turney (2001) using the Alta-Vista search engine. Applying Turney's PMI-IR technique, we found that it gave the *lowest* ratings to "Lawyers as slimeballs" (1.06) and "Lawyers as bastards" (1.15), the latter being roughly equivalent PMI-IR's rating of "Lawyers as kangaroos" (1.17)! We then asked a group of 26 undergraduates at Willamette University (Oregon) to also do these ratings. These results (Figure 1) are much more in line with one might expect a normal subcognitive profile to be for these questions — namely, lawyers are judged to be most like slimeballs, bastards, dogs and sharks, and least like telephones, kangaroos, and birds. PMI-IR, on the other hand, judges lawyers to be most like computers, cats, and telephones and least like slimeballs, bastards, kangaroos and robins. Lawyers as sharks or fish are judged to be equally bad. A comparison of human vs. PMI-IR results can be seen in Figure 1. In short, it is amply clear that the human subcognitive profile, at least for this straightforward question about lawyers, does not even vaguely resemble the PMI-IR-generated profile.

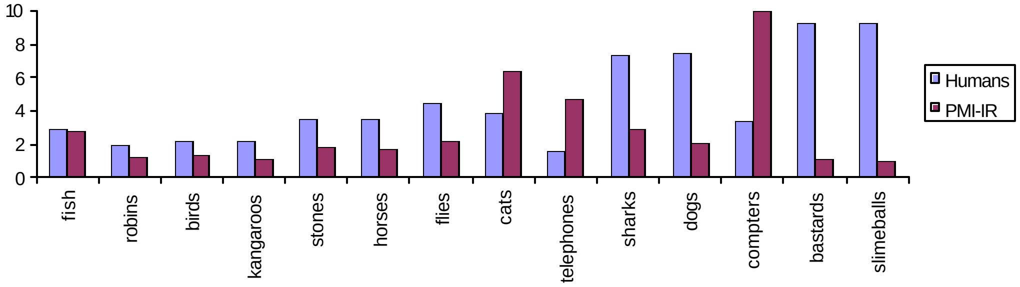

Figure 1. A comparison of PMI-IR and Human data. The two profiles are clearly very different.

We also found that PMI-IR gave an extremely high rating to "Lawyers as children," higher, in fact, than any of the choices tested in Figure 1. Clearly, something is wrong here: first, lawyers cannot even *be* children and, even metaphorically, it just doesn't seem right to us.

Rating the plausibility of the names

Next we used Turney's algorithm to judge how good various first names would be for an Israeli or a Palestinian minister. We chose ten traditional Jewish names (Uri, Ariel, Moshe, Yitzhak, Yehudi, David, Samuel, Benjamin, Shimon, and Zeev) and nine traditional Arab names (Saddam, Usama, Ahmed, Mohammed, Salah, Amin, Khalil, Ashrawi, and Yasser). We asked two separate questions, each processed independently by PMI-IR. The first was "How good is X [one of the names, e.g., *Ahmed*) as the name of an Israeli minister?" All nineteen names were rated for this question. Then a second question was asked: "How good is X [again, one of the 19 names] as the name of a Palestinian minister?" All 19 names were rated for this second question. We then compared the ratings for each name for the two questions to determine their degree of correlation.

Once again, PMI-IR fails rather spectacularly: for example, it considers *Yasser* to be almost as good a first name for an Israeli minister as for a Palestinian minister! Similarly, *Ariel* is judged to be the best name for either an Israeli minister or a Palestinian minister among all ten Jewish names as well as among all nine Arab names. The results for other names are shown in Figure 2.

Why does the program rate *Yasser* as a highly probable name for an Israeli minister and *Ariel* as highly probable for a Palestinian minister? The reason is simple: Because the program is concerned *only* with the co-occurrence of words, in this case the words *Yasser*, *Ariel*, *Israeli*, *Palestinian* and *minister*. The fact that Israel and Palestine are currently waging an undeclared war is known to PMI-IR only through higher than normal co-occurrences of war-related words and words like *Israel*, *Palestine*, *intifada*, etc. It knows nothing *about* wars, about their causes and effects, about their relations to and effects on societies and individuals in those societies, about hatred, about destruction, about refugees, about Israel, about Palestine, etc. *ad infinitum.* It knows only that sometimes these words co-occur with higher frequency than others. The complete absence in PMI-IR of this deep relational structure between the words that it encounters (and concepts these words represent) is precisely why PMI-IR fails to convincingly answer even the simplest of subcognitive questions.

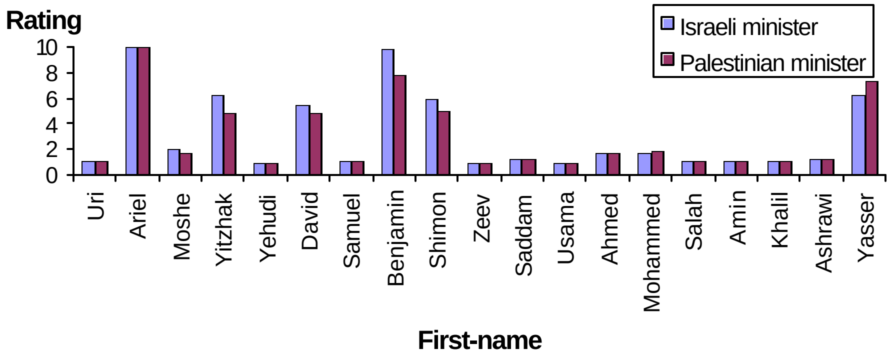

Figure 2. For the two separate questions: “How good is X as the name of an Israeli minister?” and “How good is X as the name of a Palestinian minister?” PMI-IR produces an almost a perfect correlation between the appropriateness of a given name as either that of an Israeli or a Palestinian minister.

So, to return to our example, in the context of the current crisis in the Middle East, good names for Palestinian ministers should be perceived as bad names for Israeli ministers and vice-versa. PMI-IR is, of course, unaware of the cultural context surrounding these questions. Specifically, PMI-IR is ignorant of the obvious (to us) cultural fact that some first names are typically Jewish while others are typically Arab and the relation of that cultural fact to the currently perceived inappropriateness of Palestinian ministers with Jewish names and vice-versa. So, according to PMI-IR, the appropriateness of a name for a Palestinian minister correlates almost perfectly (+0.98) with the appropriateness of the same name for an Israeli minister.

Turney might again argue that we have again just picked a special case that supports our argument. We decided to pick an example, simple in the extreme and far removed from politics and current events. We decided to compare the program’s answers to the following two questions: How good is X [a first name] as the name of a father?” and “How good is X [the same first name as in the first question] as the name of a mother?” We applied PMI-IR as described in Turney (2001) to ten very common men’s names (John, William, Stuart, Peter, Robert, Jack, Gary, Steve, Albert, and Michael) and to ten very common women’s names (Barbara, Mary, Patricia, Linda, Susan, Jennifer, Karen, Nancy, Elizabeth, and Dorothy).

When judging the appropriateness of a particular name as the name of a father (or mother), humans partly rely on a simple fact that the program does not have — namely, that fathers are invariably men, while mothers are invariably women. Consequently, humans will *necessarily* rate women’s names lower than men’s names for the question: “How good is X as the name of a father?” Not so PMI-IR. The program concludes that “John” is the best name out of all twenty names for a father *and for a mother.* It rates “Mary” very high as well as a good name for a father or for a mother. Ditto for the name “William.” As in the above example, the appropriateness of a particular name for a father correlates essentially perfectly (+0.99) with the appropriateness of that same name for a mother!

Once again, the algorithm proposed by Turney fails because extracting co-occurrences of words in a large corpus of text is simply not good enough to answer questions that require even slightly abstract contextual knowledge or experience. Again, the problem is that PMI-IR has neither abstract rules nor world experience that it can rely on. And since, in any text where the word "father" occurs, the word "mother" will generally not be far away, PMI-IR fails completely on this simple subcognitive task.

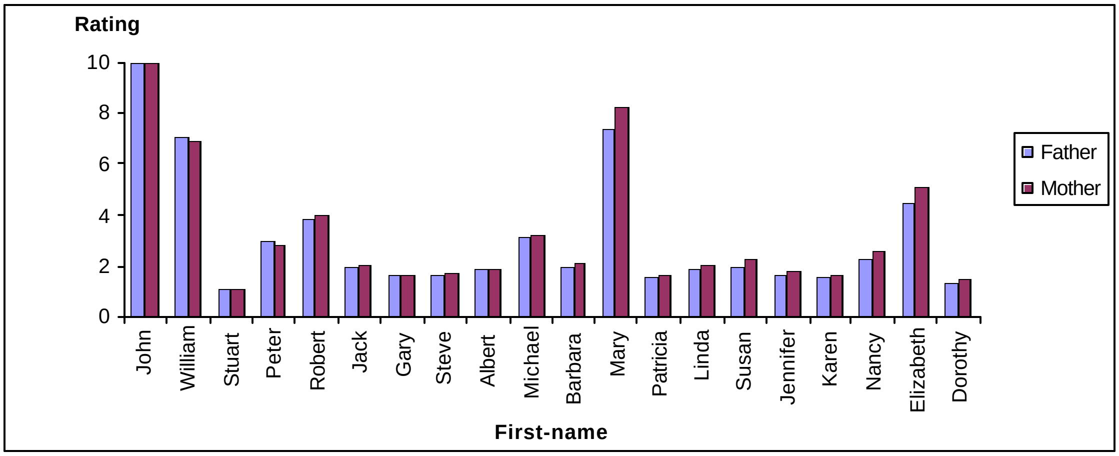

Figure 3. The two questions asked were: "How good is X as the name of a father?" and "How good is X as the name of a mother?" Once again, lacking all context about what "fathers" and "mothers" actually are, PMI-IR produces an almost a perfect correlation between the appropriateness of the names, male or female, for a father or for a mother!

Our final test of PMI-IR was having it consider the following subcognitive judgment: "Rate X as intelligent." (1 = not intelligent, 10 = extremely intelligent). As one might hope, "human beings," "men," "women," "boys" and "girls" all came out as more intelligent than "flowers" and "walls." But it also rated "cows" as more intelligent than "Americans". And further, "cows," "Americans" and "George Bush" were judged to be far less intelligent than "dogs" or "dolphins"!

A closer look at PMI-IR's "successes"

Now let us consider two cases in Turney's paper where he claims PMI-IR has succeeded. He claims qualitative success for his program for the following two subcognitive questions:

i) Rate *Flugly* as the name of an accountant in a WC Fields movie?
ii) Rate *Flugly* as the name of a glamorous Hollywood actress?

Because of a supposed problem of sparseness of proper nouns, Turney chose not to include either *Hollywood* or *W.C. Fields* in his queries when calculating PMI-IR ratings for these two questions. (Note, however, that "W.C. Fields" gets more hits than "cow/cows," five times as many as "banana/bananas" and ten times as many as "coconut/coconuts." Consequently, invoking sparseness to exclude these proper names seems, at best, somewhat strange.)

The context of "W.C. Fields" is *crucial* to our rating of *Flugly* in the question "Is *Flugly* a good name for an accountant in a W.C. Fields movie?". Our rating is based largely, if not exclusively, on how W. C. Fields would have pronounced the name *Flugly*. In our mind's ear, we hear him pronouncing the name as "Flugleeee" (which, if you have ever seen any W.C. Fields movies, is what you are doing as you read this sentence...). The question has much less to do with accountants than with characters, any characters, in a W.C. Fields movie and how their names would have been pronounced by W.C. Fields. And yet, the name of W.C. Fields is nowhere to be found in Turney's calculation of how we humans would do this particular rating. How, then, could Turney expect PMI-IR — or any program, however sophisticated — to correctly, or even qualitatively, simulate humans' judgment on this question when it lacks information crucial to making that judgment?

Now consider how Turney handled the rating of *Flugly* as the name of a glamorous Hollywood actress. Aside from the problem of dropping the proper noun "Hollywood" (which is less serious here than in the previous case), he determines this rating based only on the conditional probability Pr(Flu*|actress). Flu* means that anything at all can replace the last three letters of *Flugly* in his calculus. But a name like *Fluviana* — say, Natalya Fluviana — could be the name of a glamorous Hollywood actress. But Turney's progr am would get exactly the same results with *Fluviana* as it does with *Flugly*. Is this reasonable? No, because, one of the main reasons *Flugly* doesn't work for us is that it contains an unpleasing-to-the-ear guttural "g," to say nothing of the syllable "ug" or the entire word "ugly." (See French, 1990) But Turney simply removes the final "gly" from *Flugly* because he can't get enough hits on the Web with the whole word; he doesn't even break the word at the syllable boundary. Thus, not only does PMI-IR lack all of the contextual and semantic information available to humans, in this case, as in the preceding example, it is not even given information that is crucial in our judgment of *Flugly* as the name of a glamorous Hollywood actress.

Finally, Turney agrees that PMI-IR has problems with subcognitive questions involving "contextual information." (personal communication) But "contextual information" is precisely what makes answering subcognitive questions hard for disembodied computers and easy for us. The context is built into the questions and, either explicitly or implicitly, consists of facts about the world and about our experience with it. To be able to answer subcognitive questions in a human-like manner *requires* being able to handle context at least approximately as we humans do. We discuss this point in more detail below.

Theoretical reasons for PMI-IR's failure .

The major theoretical problem with PMI-IR is this. It has no semantics or any experience with the world allowing it to correctly situate the utterances that it is asked to judge. It doesn't know about wars or their effects, or that mothers are female and fathers are male, etc. and it has no way of learning this, either through experience or by having been explicitly programmed with this knowledge.

Turney seems to believe that because PMI-IR does well in selecting the correct synonym for a given word from a short list of possibilities, that the same program can be used to answer subcognitive questions correctly. Let us briefly examine the differences that are involved in the two tasks.

By searching hundreds of millions of Web pages, PMI-IR can do better on a synonym test than any other current computer program. This is believable and reasonable. Turney illustrates PMI-IR's performance on the synonym-finding task with the word *levy* (as in "to levy taxes"). Four choices are proposed — *imposed, believed, requested, correlated* — and the program chooses one of them as the best synonym based on how often that word is close to "levy" in many Web pages. The reason for PMI-IR's success on this task is not hard to see. It involves the stylistic reasons for which we use synonyms — viz., so as not to repeat the same word too often in a given text or, especially, in the same paragraph. This constraint imposes the proximity of synonyms, which is detected by PMI-IR.

Assume you are writing a Web article about some *blunder* that occurred. In describing this blunder, you are aware that it is bad style to repeat the word *blunder* over and over again in your text, so you resort to synonyms, such as *failure*, *mishap*, *mistake*, *slip*, *bungle*, *mess*, and so on. This obviously produces co-occurrences of *blunder* and *mistake*, of *blunder* and *slip*, etc., and this is precisely what PMI-IR detects. A *blunder* IS (to a first approximation) a *mistake*, which IS a slip, etc. Let us call this *attributional similarity*. We can expect attributionally similar words, if only for stylistic reasons, to occur close to one another in a text. Hence, PMI-IR's excellent performance on this task.

On the other hand, answering subcognitive questions requires a great deal than this. Consider rating a *banana split* as *medicine*. The number of times that these two items will occur together in a text anywhere on the Internet is now and will forever be infinitesimally small compared to the other associations involving banana splits or medicine. Turney tends to dismiss this problem as being a problem of sparseness. But it cannot be dismissed; this issue is at the very heart of the why PMI-IR will never be adequate. *Of course* the number of Web pages containing both terms will be vanishing small because is it not a common association at all, but it remains a perfectly valid one and one that we can judge without difficulty because we understand it both *in relation to our experience with the world*, i.e., with facts like the doctor bringing us a bowl of ice-cream after you had our tonsils out, with our mother taking us for a sundae to pick up our spirits when our junior high school safety poster was eliminated from the city competition, etc.

In other words, describing one word in terms of another usually involves much more than the above kind of "blunder-mistake-mishap-slip" synonym searching. It involves mentally placing the both words in a variety of *relational* as well as attributional contexts (that can shift fluidly) and converging on a context that fits both words (for detailed discussions of this see: Chalmers, French, & Hofstadter, 1992; Mitchell, 1993; Hofstadter, 1995; French, 1995, etc.) If both words fit that context very well, then we give the association a high rating. The more difficult it is to converge on an appropriate context for both words, the lower the rating.

PMI-IR, however, is incapable of extracting these all-important relational and contextual characteristics of situations. Specifically, in the case of subcognitive questions of the form, "Rate X as a Y," the program is incapable of grasping the relation of the concepts in the query with the rest of human experience, the ability that allows us humans to judge the degree to which X is "the same as" Y.

Even if PMI-IR could correctly answer subcognitive questions...

Let us assume for a moment that PMI-IR were able to answer some range of subcognitive questions correctly. Would this be sufficient to rebut French's (1990) argument that subcognitive questions could always be used to unmask a computer that had not experienced the world in a human manner? No, unless we assume that PMI-IR could answer *all* subcognitive questions in a manner that was indistinguishable from humans.

Remember that our Turing Test interrogator is a *very* smart woman, thoroughly up to date on the latest techniques of artificial intelligence, psychology, etc. Consequently, she would know all about PMI-IR. So, once she had drawn up her list of subcognitive questions and tested a large random sample of people on them, allowing her to draw up her Subcognitive Profile, she would then submit all the questions to PMI-IR. The questions that PMI-IR answered in a human-like manner *would then be eliminated from her list*. Only those questions that were explicitly not able to be answered by co-occurrence calculations would be retained. And, as is clear from the empirical section of this paper, a lot of questions would still remain on her list to make up the Subcognitive Profile. Only when there are no questions left on her Subcognitive Profile can we say that, indeed, we can no longer unmask the computer with subcognitive questions. But at present we are still a very long way from this goal and it is doubtful that techniques that rely on co-occurrence matrices for disembodied computers will ever get us there.

Conclusions

We have attempted to show that correctly answering subcognitive questions (French, 1990, 2000a) is considerably harder than Turney (2001) suspects and cannot be done with any reliability using the simple algorithm that he proposes. Further, we have only dealt with a single type of subcognitive question in the present paper. It turns out that subcognitive questions can also be used to probe even the physical level (French, 2000a), which means that human embodiment (or a computer's non-embodiment) can be directly tested via subcognitive questions. The quirky human way in which our muscles are connected, the lengths of our fingers, the extent of our arms, the position of our eyes and the "illogical" way in which they are connected to the brain, etc. can be probed by careful subcognitive questions that require simple corporal self-experiments in order to be answered. We see no way that a program like PMI-IR could ever come close to answering questions like these.

The issues raised in this article apply, of course, to the limits of "intelligent" search algorithms used on large corpora that extract co-occurrences matrices. Their lack of semantics and their inability to correctly and reliably contextualize pose serious, if not insurmountable, problems for these methods. The hope of researchers in this area is that semantics and context will either ultimately be unnecessary or will emerge from ever more sophisticated manipulations of co-occurrence matrices. In the few examples we have included in this article, we can see just how difficult — and most likely, impossible — the co-occurrence road will be for these researchers to travel.

In short, for empirical as well as theoretical reasons, we believe that Turney is premature in his dismissal of the power of subcognitive questions in a Turing Test to unmask a computer that had not lived life as we humans had.

Acknowledgments

This work was supported in part by grant HPRN-CT-1999-00065 from the European Commission. Christophe Labiouse is supported by a Belgian NFSR Research Fellowship. The authors would also like to thank Jim Friedrich at Willamette University, Oregon, for his help in conducting the informal survey cited in the text.

References

Chalmers, D. J., French, R. M. and Hofstadter, D. R. (1992). High-level Perception, Representation, and Analogy: A Critique of Artificial Intelligence Methodology.

Journal of Experimental and Theoretical and Artificial Intelligence, *4*(3), 185-211.

Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1-32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.), *Selected Papers of J.R. Firth 1952-1959*, London: Longman (1968).

French, R. M. (1990). Subcognition and the Limits of the Turing Test. *Mind*, *99*(393), 53-65.

French, R. M. (1995). *The Subtlety of Sameness: A theory and computer model of analogy-making*. Cambridge, MA: MIT Press.

French, R. M. (2000a). Peeking Behind the Screen: The Unsuspected Power of the Standard Turing Test. *Journal of Experimental and Theoretical Artificial Intelligence, 12*, 331-340.

French, R. M. (2000b). The Turing Test: the first 50 years. *Trends in Cognitive Sciences*, *4*(3), 115-121.

Hofstadter, D. R. and the Fluid Analogies Research Group (1995). *Fluid Concepts and Creative Analogies*, New York, NY: Basic Books.

Mitchell, M. (1993) *Analogy-making as Perception: A computer model*. Cambridge, MA: MIT Press.

Turney, P. (2001). Answering Subcognitive Turing Test Questions: A Reply to French. *Journal of Experimental and Theoretical Artificial Intelligence*. (see this issue).