Seminar paper (B-uppsats), Spring 2010
Philosophy, Umeå University

Author: Anders Sundman
Supervisor: Alexander Almér

The Meaning of Spam

Can Computers Represent Semantic Contents?

ABSTRACT

In this paper I argue that the the Chinese room thought experiment, invented by John R. Searle, does not show that a computer can not represent semantic contents. By analogy with a spam filter, I use Putnam's 4-tuple normal form definition of meaning to show that a spam filter actually does represent meaning (semantic contents). While this paper advances arguments for a view that computers can represent semantics as well as syntax; I make no claims that computers can understand anything. On the contrary, I argue that syntax and semantics is not sufficient for understanding.

Introduction

In his 1983 article *Can Computers Think?* John R. Searle introduced the now famous Chinese room argument. This argument is intended to show that computers are only capable of syntactic symbol manipulation, unlike humans who are also capable of representing semantically meaningful mental content. Searle's conclusion is that, since computers lack this ability, they can not think.

I agree with Searl in that I do not think that computers can think. However, I don't share his reasons for this conclusion. Contrary to Searle I think that computers can, and in fact already do, represent meaningful semantic content. Instead I will argue that there is something else, a phenomenological experience and a conscious awareness, that a general computer can not have and that is an integral part of thinking.

In this paper I present the email spam filter as a variation on Searle's Chinese room. This example will show that a computer program, the spam filter, does know the meaning of spam. It's not going to be the same meaning of spam email as you and I have, but it's going to be *a* meaning. The paper will further show that semantics (representing meaning) is not sufficient for *understanding* what spam is. Agreeing with Searl, I conclude by arguing that *understanding* requires something deeper than syntax and semantics.

Background

Before diving into the investigation of how a computer program might actually represent meaning, I would like to provide some cursory background information about the problem domain. In the following three sections we will see a brief overview of Searle's argument known as the Chinese room, I will introduce the concept of meaning more formally and we will investigate how a spam filters works.

The Chinese Room

Lets start by having a closer look at what claims it is that Searl is actually making. In his example, a person is locked inside a room together with a rule book and a basket full of Chinese symbols. When a Chinese symbol is passed into room, the rule book specifies what symbol should be picked from the basket and passed out of the room.

Unknown to the person inside the room, the symbols coming in are questions and the symbols passed out are answers. If the rule book is really good, so Searl claims, the person outside the room, a native Chinese speaker, could not determine if the person inside the room understands Chinese or not.[1]

But, and this is Searle's main point, the person inside the room does not understand Chinese and he could never learn to understand Chinese in this way because he is only performing syntactic symbol manipulation. This is of course exactly what computers do as well, hence computers can not understand Chinese no matter how complex programs we might write to simulate the behavior of someone who does have a genuine understanding of the language.

[1] David J. Chalmers (ed.), *Philosophy of mind : Classical and Contemporary Readings*, New York, Oxford University Press, 2002, p. 671.

What is Meaning?

For Searl, representing semantic content, having meaning attached to symbols, is a necessary condition for having a mind. And according to him, no amount of syntactical symbol manipulation can create meaning.

So what is meaning and semantic content supposed to be? We need to clarify this before we can hope to reach any conclusions about whether computers have it or not.

Meaning, like some other philosophical concepts such as truth, are intellectual mine fields. Deceptively easy to grasp, because we have a commonsensical understanding of them, they prove elusive and mind numbingly complex when one looks a bit closer. I do not intend to perform any in depth investigations of the concept of meaning in this paper; rather I shall use the results of an analysis performed by Hilary Putnam in his 1975 article *The Meaning of 'Meaning'*.

In his article, Putnam acknowledges that meaning has historically been considered an ambiguous concept. On the one hand, meaning has been used to describe the *extension* of a term; the extension is the set of things that the term can be applied to. On the other hand, there has to be something more to meaning than extension, this something is called "intension" by Putnam.[2]

It is quite straight forward to argue that meaning is more than extension. The old Fregian example of morning star and evening star shows this. The two terms have the same extension, the planet Venus as it were, but different meanings.[3]

Through an example of a parallel world, a twin earth, Putnam shows that psychological states can not be identified with meaning. A person on earth and another person on the twin earth could possibly have the same psychological state, the same concept, but have different meanings attached to them due to a difference in extension. He summarizes his standpoint like this:

> *"... meaning cannot be identified with extension. Yet it cannot be identified with 'intension' either, if intension is something like an individual speaker's concept."* [4]

In conclusion, Putnam advances another analysis of the meaning of 'meaning'. He suggest that the meaning of a term has four components: *syntactic markers*, *semantic markers*, *stereotype* and *extension*. This is a schema, a normal form, for representing and describing meaning. He writes:

> *"If we know what a 'normal form description' of the meaning of a word should be, then, as far as I am concerned, we know what meaning is in any scientifically interesting sense."* [5]

We will return to the normal form description later, but for now it's enough to note that meaning is something that can be given a 4-tuple description.

What is a Computer?

Before we can ask whether a computer can represent meaning or not, we have to clarify what a computer is more formally. From remarks about a tape, printing and erasing symbols, etc.[6] we can

[2] Chalmers (ed.), *Philosophy of mind*, p. 581.

[3] Daniel Birnbaum and Sven-Olov Wallenstein, *Gottlob Frege: Skrifter i urval*, Stockholm, Thales, 1995, p. 35.

[4] Chalmers (ed.), *Philosophy of mind*, p. 593.

[5] Ibid., p. 594.

[6] Ibid., p. 670.

guess that it is a Turing Machine (TM) of some sort that Searle has in mind in his article. Further remarks about being able to run an infinite number of programs on the same hardware indicates that Searl is actually writing about a Universal Turing Machine (UTM).

A UTM is an abstract mathematical construct. It is a TM that can simulate any other TM with any desired input. I'm not going to concern myself with the ontology of mathematical objects here; it doesn't matter what type of existence we ascribe to entities like numbers or TMs. The debate is not whether an abstract UTM can represent semantic content or not, but if a realization of a UTM can do that.

A realization of a UTM is a physical object with certain defining properties. It can execute a number of actions and update a state. There are many examples of realized UTMs, in fact all PCs are UTMs and all humans can learn to be UTMs by learning to follow some simple rules and perform some actions.

Searls argument aims to show that having the properties that makes something a UTM is not sufficient for having a capacity for representing semantic content.

A Spam Filter

Unsolicited bulk email, junk mail, intrusive advertising, offers to enlarge male genitalia, spam is a pestilence on the digital world. What a great fortune then that we have email spam filters (SFs). These helpful programs scan incoming emails and classify them as spam or not-spam. But how can a computer program, the SF know if an email is spam? Does it understand the meaning of spam?

Let me start by giving a description of a typical SF. Any modern SF is of course a very complex computer program, and a detailed description is beyond the scope of this paper. The descriptions I offer will be simplified and idealized, but not too far from how SFs actually work.

Since the late 1990s, most SFs have been using (at least) a technique known as Bayesian filtering.[7] This is a statistical method that processes each word or phrase in an email message and estimates the likelihood that the email is a spam message based on the fact that it contains this particular word. The "spam-factor" of every word is then combined (e.g. averaged) to reach some conclusion about the message as a whole.

The really clever part with this method is that if the program makes a mistake, the user can indicate this to the program and the program will update the word probabilities for the words in the email, and thus be less likely to make the same mistake again.

Let me illustrate this with an example: Mr. Resale works in advertising. He routinely receives (and sends) email containing words like 'discount', 'offer', etc.. Most spam filters would classify Mr. Resales emails as spam because there is (in general) a high probability that mails containing words like 'offer' are spam emails. However, after using the SF for a while, moving the legitimate emails from the spam folder to the in box, Mr. Resale would have caused the SF to update it's internal probabilities; and the SF would no longer consider the presence of the word 'offer' as an indication of that the email is a spam message.

[7] M. Sahami, et al., *A Bayesian Approach to Filtering Junk E-mail*, AAAI Workshop on Learning for Text Categorization (Technical Report WS-98-05), California, AAAI Press, 1998.

A bit more schematically, the spam filtering can be described in the following way:

An email arrives, it is represented as a sequence of symbols in the computers memory. The email is passed on to the SF-algorithm. The spam-word probabilities are also represented in the computer memory. A sequence of program steps are carried out that processes the input (email) symbols according to some fixed rules (using the spam-word probabilities). The formal process attaches a "is spam" or "is legitimate" flag to the email; the labels are also represented by some sequence of symbols in the computers memory.

The human user then gives feedback to the spam filter by accepting or changing the spam-labels of the email. When this happens the SF update routine kicks in, again working according to some fixed set of rules, and the symbols in its memory representing spam-word probabilities are modified.

Notice the similarity to Searl's Chinese room. Symbols come in (say, chinese emails) and symbols are sent out (say, chinese spam-labels). The entire process is governed by a fixed and rigid rule book (the program). The only real difference in the spam filter case is that spam filter contains an internal stack of symbols representing the spam-word probabilities and that the output symbols are causally connected, through the human user, with some special input symbols that the rule-book dictates should update the internal spam-word symbols.

Investigation

Now that we have a basic understanding of how the spam filter works, we can move on to the more interesting question; does the spam filter know what spam is? Does it have semantic content?

Searle's Argument

Lets start by taking a closer look at Searle's Chinese room argument. For the sake of clarity, we need to summarize what it is that Searl is saying a bit more formally. He writes:

> *"There is more to having a mind than having formal or syntactical processes. Our internal mental states, by definition, have certain sorts of content. [...] my mental state has a certain mental content in addition to whatever formal features it might have. That is, even if my thoughts occur to me in strings of symbols, there must be more to the thought than the abstract strings, because strings by themselves can't have any meaning."* [8]

A bit later Searle writes these two passages:

> *"Understanding a language, or indeed, having mental states at all, involves more than just having a bunch of formal symbols. It involves having an interpretation, or a meaning, attached to those symbols."* [9]

> *"My thoughts, and beliefs, and desires are about something, or they refer to something, or they concern states of affairs in the world; and they do that because their content directs them at the state of affairs in the world."* [10]

[8] Chalmers (ed.), *Philosophy of mind*, p. 670.

[9] Ibid., p. 671.

[10] Ibid., p. 674.

From the excerpts, and from his punch line "a computer has syntax, but no semantics"[11], we can conclude that Searl makes the following claims (among others):

i) Semantic content is meaning attached to a symbol. Strings of symbols, by themselves, do not have meaning.

ii) Human mental states have mental content.

iii) Human mental content has meaning because its representation (the thoughts) refer to something in the world.

iv) Understanding a symbol involves having a meaning attached to the symbol.

v) Computers can execute syntactical processes, but not represent semantic content.

The key points of his argument seem to be that a human, having an instance of a particular mental content, can be said to understand a symbol because the thought that represents the mental content refers to something in the world *(ii)* and *(iii)*. In this view, Searl equates understanding with representing meaning *(iv)*. He also emphasizes that meaning (semantic content) must come from a relationship between a symbol and the world *(i)* and *(iii)*.

Searle was a distinguished philosopher and it is not surprising that he anticipated objections similar to the one I'm making. In his original article he responds to an objection that he calls "the robots reply". The argument goes like this. We connect a video camera to a robot and translate the video signal to Chinese symbols. Then we put the Chinese room inside the robot (it would have to be pretty big, but never mind) and feed the Chinese video symbols into the room. The symbols coming out of the room are Chinese symbols that the robot uses to control its moving about in the world. Wouldn't the person pushing symbols inside the room *then* be able to learn something about the world, to represent semantic content? No, says Searle. The situation has not changed in any way for the person inside the Chinese room; putting it in a robot doesn't change the fact that the person in the room is just syntactically manipulating symbols.[12]

Searle's dismissal of the robots reply is just as simple and ingenious as his original thought experiment. In my opinion it is however too broad. Just like in the original example, Searle equates representing meaning with understanding and goes on to show that computers can't do that. The response to the robots reply restates this dogma and adds that the causal origins of the syntactical symbols doesn't change anything. I fully agree with Searl that his argument and response to the robots reply show that a computer can not understand any concepts. But since I differentiate between *understanding* and *representing meaning*, it doesn't immediately follow that computers can't represent meaning. If humans represent meaning because their thoughts are about the world *(iii)*, why then should it not be possible for computers to represent meaning if their internal representations (e.g. spam word probabilities) are about the world?

Semantics Demystified

Lets now return to Putnam's normal form description of meaning for a closer look. In Searle's terminology semantic content and meaning are used interchangeably, so the normal form is equally applicable to semantic content.

[11] Ibid., p. 671.

[12] Chalmers (ed.), *Philosophy of mind*, p. 672.

There are four components to the normal form according to Putnam.

1. Syntactic Markers.

   Is the term a noun or a verb? How is it combined with other terms according to the rules of grammar to create sentences.

2. Semantic Markers.

   What relation does the term have to other terms or to regularities in the world? If the term is syntactically a kind-name, does it name a kind of animal or a kind of mineral? The *semantic marker* should not be confused with the term *semantic content* as used Searle. In both cases however, it is about a relationship between a concept and the world.

3. Stereotype.

   A representative image of the term. A list of typical properties. The stereotype is similar to the old Platonic concept of *idea*, but without the ontological and metaphysical claims.

4. Extension.

   This is the actual extension of the term. Putnam points out that we typically don't have perfect access to, and knowledge of, how to specify the extension. In the case of the term *water*, the extension is all water in the universe. In antiquity water could be described as a clear liquid. That we only recently have learned that water can be accurately described as the substance consisting of molecules with two hydrogen atoms and one oxygen atom does not change the content of the extension. The extension of water, from antiquity till today, has always been all the water there is.

This 4-tuple is all there is to meaning according to Putnam.[13] If we have all four components, there is nothing missing as far as meaning is concerned. Representing the four components is necessary and sufficient for representing meaning. To show that the spam filter can represent meaning, we have to show that it can represent the four components.

So, lets review the components again; this time testing how the spam filter fares in each category. Remember that the filter does not have to have the same meaning of spam as a human, it's enough that it has *some*, possibly very limited meaning.

1. Syntactic Markers.

   In the context of computers programs, what does syntax mean? It might be a bit awkward to think of computer programs, with their input and output, to be a language proper. But its not as strange as it might first sound. Think of it as a language game á la Wittgenstein[14]. There is of course the internal syntax of the computer program, the sequence of symbols that determines it's operation, but that is not what we are concerned about here. Instead think of the computer as a black box (or a black room). The rules of the spam filter language, its grammar, is very simple. Messages are passed to the program and the program "answers" with "it's spam" or "it's not spam". Anyone, or any program that can play this language game (never mind the content of the message) has a well defined syntactic "spam marker". This

[13] Chalmers (ed.), *Philosophy of mind*, p. 594.

[14] Ludwig Wittgenstein, *Philosophical Investigations*, 3rd Ed., translated by G. E. M. Anscombe, Padstow (United Kingdom), Blackwell Publishing, 2001.

might strike you as embarrassingly trivial, but notice that many program, like word processors or MP3 players, can not play the "spam language game" and do not have the syntactic markers of the spam concept.

2. Semantic Markers.

   The syntactic markers just says that the computer program can play the language game according to the grammar. It doesn't even have to consider the contents of the messages. A well trained spam filter will however not respond with "is spam" or "isn't spam" in a random fashion. It's responses will depend on the content of the message as well as on the experience of prior interaction with the user. This adds something beyond the formal syntactical competence to the language game; this something is a semantic marker. Notice here that we are not saying anything about computer programs executing programs by syntactical manipulation. The computer is still a black box and we really don't care about how its innards work. The important thing is that the programs actions are guided by prior interaction with the world. The spam / not-spam flags are about the world; they are about how the user has sent feedback signals to the computer program.

3. Stereotype.

   The spam filter has an internal representation of spam word probabilities. We could easily add a function to the spam filter that would, on request give us a list of the most prominent features of a spam message; the top ranking words. We could also add a function that would make the spam filter generate a typical spam message by combining high ranking words according to some natural language production rules. The spam filter can thus be said to represent stereotypical information.

4. Extension.

   This is the set of all messages that the spam filter would, at the current time, classify as spam. To find out if a message is in the extension, just run it through the filter.

And there you have it! From the explication above we see that spam for the spam filter does in fact have meaning. The syntactic markers are given by the programs design. The semantic markers are given by program design in combination with interaction with the world (the user). The stereotype is represented as spam word probabilities. The extension (or at least a part of it) is what you find in your email programs "spam" folder.

Syntax and Semantics, so what?

So far I have tried to show that some concepts that computers represent *can* have meaning. Does this mean that computers *understand* concepts like spam?

I don't think so. In fact, I agree with Searle's claim that *understanding* a symbol involves having a *meaning* attached to the symbol. But unlike him, I don't think this is a sufficient condition.

Like Searle I think that brains cause minds and that no functional analysis can accurately describe consciousness and phenomenological experience. In particular I think that *understanding* is a conscious process and that the phenomenological experience of this process is central to the activity of understanding. Our understanding of the concept 'understanding' involves more than it's meaning, it involves a conscious and phenomenological experience.

Although syntax and semantics can be handled by computers, understanding can not; simply because the properties that makes something a computer are not sufficient to cause a conscious mind.

Conclusions

Saying that computers only manipulate symbols is like saying that a book is only a collection of letters. It's true, but it's beside the point. A book can be about something just because it is collection of letters and computers can represent meaning (have semantic content) just because they manipulate symbols; it's the mechanism they use to get the job done.

Using the definition of 'meaning' proposed by Putnam, I have shown how a spam filter can acquire and represent meaning. spam filters are computer programs. So by example, a computer can represent the meaning of a concept, because the properties that makes something a computer are sufficient for this task. The same properties do however neither cause consciousness nor phenomenological experiences and are hence not sufficient for understanding.

References

David J. Chalmers (ed.), *Philosophy of mind : Classical and Contemporary Readings*, New York, Oxford University Press, 2002.

M. Sahami, S. Dumais, D. Heckerman, E. Horvitz, *A Bayesian Approach to Filtering Junk E-mail*, AAAI Workshop on Learning for Text Categorization (Technical Report WS-98-05), California, AAAI Press, 1998.

Ludwig Wittgenstein, *Philosophical Investigations* (3rd Ed)., translated by G. E. M. Anscombe, Padstow (United Kingdom), Blackwell Publishing, 2001.

Daniel Birnbaum and Sven-Olov Wallenstein, *Gottlob Frege: Skrifter i urval (Gottlob Frege, Über Sinn und Bedeutung, 1892)*, Stockholm, Thales 1995.