The Winograd Schema Challenge

Hector J. Levesque
Dept. of Computer Science
University of Toronto
Toronto, Ontario
Canada M5S 3A6
hector@cs.toronto.edu

Abstract

In this paper, we present an alternative to the Turing Test that has some conceptual and practical advantages. Like the original, it involves responding to typed English sentences, and English-speaking adults will have no difficulty with it. Unlike the original, the subject is not required to engage in a conversation and fool an interrogator into believing she is dealing with a person. Moreover, the test is arranged in such a way that having full access to a large corpus of English text might not help much. Finally, the interrogator or a third party will be able to decide unambiguously after a few minutes whether or not a subject has passed the test.

Introduction

The well-known Turing Test was first proposed by Alan Turing (1950) as a practical way to defuse what seemed to him to be a pointless argument about whether or not machines could think. In a nutshell, he proposes that instead of asking such a vague question and then getting caught up in a debate about what it means to really be thinking, we should focus on *observable behaviour* and ask whether a machine would be capable of producing behaviour that we would say required thought in people. The sort of behaviour he had in mind was participating in a natural conversation in English over a teletype in what he calls the Imitation Game. The idea, roughly, is that if an interrogator were unable to tell after a long, free flowing and unrestricted conversation with a machine whether she was dealing with a person or a machine, then we should be prepared to say that the machine was thinking. Requiring more of the machine, such that as that it look a certain way, or be biological, or have a certain causal history, is just arbitrary chauvinism.

It is not our intent to defend Turing's argument here (but see the Discussion section below). For our purposes, we simply accept the argument and the emphasis Turing places on intelligent behaviour, counter to critics such as Searle (1980). We also accept that typed English text is a sufficient medium for displaying intelligent behaviour, counter to critics such as Harnad (1989). That is, assuming that *any* sort behaviour is going to be judged sufficient for showing the presence of thinking (or understanding, or intelligence, or whatever appropriate mental attribute), we assume that typed English text, despite its limitations, will be a rich enough medium.

The trouble with Turing

The Turing Test does have some troubling aspects, however. First, note the central role of *deception*. Consider the case of a future intelligent machine trying to pass the test. It must converse with an interrogator and not just show its stuff, but fool her into thinking she is dealing with a *person*. This is just a game, of course, so it's not really lying. But to imitate a person well without being evasive, the machine will need to assume a false identity (to answer "How tall are you?" or "Tell me about your parents."). All other things being equal, we should much prefer a test that did not depend on chicanery of this sort. Or to put it differently, a machine should be able to show us that it is thinking without having to pretend to be somebody or to have some property (like being tall) that it does not have.

We might also question whether a *conversation* in English is the right sort of test. Free form conversations are no doubt the best way to get to know someone, to find out what they think about something, and therefore *that* they are thinking about something. But conversations are so adaptable and can be so wide-ranging, they facilitate deception and trickery.

Consider, for example, ELIZA (Weizenbaum 1966), where a program (usually included as part of the normal Emacs distribution) using very simple means, was able to fool some people into believing they were conversing with a psychiatrist. The deception works at least in part because we are extremely forgiving in terms of what we will accept as legitimate conversation. A Rogerian psychiatrist may say very little except to encourage a patient to keep on talking, but it may be enough, at least for a while.

Consider also the Loebner competition (Shieber 1994), a restricted version of the Turing Test that has attracted considerable publicity. In this case, we have a more balanced conversation taking place than with ELIZA. What is striking about transcripts of these conversations is the fluidity of the responses from the subjects: elaborate wordplay, puns, jokes, quotations, clever asides, emotional outbursts, points of order. Everything, it would seem, except clear and direct answers to questions. And how is an interrogator supposed to deal with this evasiveness and determine whether or not there is any real comprehension behind the verbal acrobat-

ics? More conversation. "I'd like to get back to what you said earlier." Unsurprisingly, short conversations are usually inconclusive, and even with very long ones, two interrogators looking at the same transcript may disagree on the final verdict. Grading the test, in other words, is problematic.

How can we steer research in a more constructive direction, away from deception and trickery? One possibility is something like the *captcha* (von Ahn *et al* 2003). The idea here is that a distorted image of a multidigit number is presented to a subject who is then required to identify the number. People in general can easily pass the test in seconds, but current computer programs have quite a hard a time of it (cheating aside).[1] So this test does, at least for now, distinguish people from machines very well. The question is whether this test could play the role of the Turing Test. Passing the test clearly involves some form of cognitive activity in people, but it is doubtful whether it is thinking in the full-bodied sense that Turing had in mind, the touchstone of human-level intelligence. We can imagine a sophisticated automated digit classifier, perhaps one that has learned from an enormous database of distorted digits, doing as well as people on the test. The behaviour of the program may be ideal; but the scope of what we are asking it to do may be too limited to draw a general conclusion.

Recognizing Textual Entailment

In general, what we are after is a new type of Turing Test that has these desirable features:

- it involves the subject responding to a broad range of English sentences;
- native English-speaking adults can pass it easily;
- it can be administered and graded without expert judges;
- no less than with the original Turing Test, when people pass the test, we would say they were thinking.

One promising proposal is the *recognizing textual entailment* (RTE) challenge (Dagan *et al* 2006; Bobrow *et al* 2007; Rus *et al* 2007). In this case, a subject is presented with a series of yes-no questions concerning whether one English sentence (A) *entails* another (B). Two example pairs adapted from (Dagan *et al* 2006) illustrate the form:

- A: Time Warner is the world's largest media and internet company.
  B: Time Warner is the world's largest company.
- A: Norway's most famous painting, "The Scream" by Edvard Munch, was recovered Saturday.
  B: Edvard Munch painted "The Scream."

This is on the right track, in my opinion. Getting the correct answers (no and yes above, respectively), clearly requires some thought. Moreover, like the captcha, but unlike the Turing Test, an evasive subject cannot hide behind verbal maneuvers. Also, in terms of a research challenge, incremental progress on the RTE is possible: we can begin with simple lexical analyses of the words in the sentences, and then progress all the way to applying arbitrary amounts of world knowledge to the task.

A problem with this challenge, however, is that it rests on the notion of entailment. Of course there is a precise definition of this concept (assuming a precise semantics, like in logic), but subjects would not be expected to know or even understand it. The researchers instead explain to subjects that "T entails H if, typically, a human reading T would infer that H is most likely true" (Dagan *et al* 2006). The fact that we need to predict what humans would do, and the use of "typically" and "likely" are troubling. What if the second (B) above was this:

- B: The recovered painting was worth more than $1000.

Technically, this is not an entailment of (A), although it would certainly be judged true! Of course, subjects can be trained in advance to help sort out issues like this, but it would still be preferable for a practical test not to depend on such a delicate logical concern.

What we propose in this paper is a variant of the RTE that we call the *Winograd Schema* (or WS) challenge. It requires subjects to answer binary questions, but without depending on an explicit notion of entailment.

The Winograd Schema Challenge

A WS is a small reading comprehension test involving a single binary question. Two examples will illustrate:

- The trophy would not fit in the brown suitcase because it was too big. What was too big?

  Answer 0: the trophy
  Answer 1: the suitcase

- Joan made sure to thank Susan for all the help she had given. Who had given the help?

  Answer 0: Joan
  Answer 1: Susan

We take it that the correct answers here are obvious. In each of the questions, we have the following four features:

1. Two parties are mentioned in a sentence by noun phrases. They can be two males, two females, two inanimate objects or two groups of people or objects.
2. A pronoun or possessive adjective is used in the sentence in reference to one of the parties, but is also of the right sort for the second party. In the case of males, it is "he/him/his"; for females, it is "she/her/her" for inanimate object it is "it/it/its," and for groups it is "they/them/their."
3. The question involves determining the referent of the pronoun or possessive adjective. Answer 0 is always the first party mentioned in the sentence (but repeated from the sentence for clarity), and Answer 1 is the second party.
4. There is a word (called the *special* word) that appears in the sentence and possibly the question. When it is replaced by another word (called the *alternate* word), everything still makes perfect sense, but the answer changes.

[1] Cheating will always be a problem. The story with captchas is that one program was able to decode them by presenting them on a web page as a puzzle to be solved by unwitting third parties before they could gain access to a free porn site! Any test, including anything we propose here, needs to be administered in a controlled setting to be informative.

We will explain the fourth feature in a moment. But note that like the RTE there are no limitations on what the sentences can be about, or what additional noun phrases or pronouns they can include. Ideally, the vocabulary would be restricted enough that even a child would be able to answer the question, like in the two examples above. (We will return to this point in the Incremental Progress section below.)

Perhaps the hardest item to justify even informally from the requirements in the previous section is that *thinking* is required to get a correct answer with high probability. Although verbal dodges are not possible like in the original Turing Test, how do we know that there is not some trick that a programmer could exploit, for example, the word order in the sentence or the choice of vocabulary, or some other subtle feature of English expressions? Might there not be some unintended bias in the way the questions are formulated that could help a program answer without any comprehension?

This is where the fourth requirement comes in. In the first example, the special word is "big" and its alternate is "small;" and in the second example, the special word is "given" and its alternate is "received." These alternate words only show up in alternate versions of the two questions:

- The trophy would not fit in the brown suitcase because it was too small. What was too small?

  Answer 0: the trophy
  Answer 1: the suitcase

- Joan made sure to thank Susan for all the help she had received. Who had received the help?

  Answer 0: Joan
  Answer 1: Susan

With this fourth feature, we can see that clever tricks involving word order or other features of words or groups of words will not work. Contexts where "give" can appear are statistically quite similar to those where "receive" can appear, and yet the answer must change. This helps make the test *Google-proof:* having access to a large corpus of English text would likely not help much (assuming, that answers to the questions have not yet been posted on the Web, that is)! The claim is that doing better than guessing requires subjects to figure out what is going on: for example, a failure to fit is caused by one of the objects being too big and the other being too small, and they determine which is which.

The need for thinking is perhaps even more evident in a much more difficult example, a variant of which was first presented by Terry Winograd (1972), for whom we have named the schema:[2]

The town councillors refused to give the angry demonstrators a permit because they feared violence.
Who feared violence?

Answer 0: the town councillors
Answer 1: the angry demonstrators

Here the special word is "feared" and its alternate is "advocated" as in the following:

The town councillors refused to give the angry demonstrators a permit because they advocated violence.
Who advocated violence?

Answer 0: the town councillors
Answer 1: the angry demonstrators

It is wildly implausible that there would be statistical or other properties of the special word or its alternate that would allow us to flip from one answer to the other in this case. This was the whole point of Winograd's example! You need to have background *knowledge* that is not expressed in the words of the sentence to be able to sort out what is going on and decide that it is one group that might be fearful and the other group that might be violent. And it is precisely bringing this background knowledge to bear that we informally call *thinking*. The fact that we are normally not *aware* of the thinking we are doing in figuring this out should not mislead us; using what we know is the only explanation that makes sense of our ability to answer here.

A library in standard format

In constructing a WS, it is critical to find a *pair* of questions that differ in one word and satisfy the four criteria above. In building a library of suitable questions, it is convenient therefore to assemble them in a format that lists both the special word and its alternate. Here is the first example above in this format:

The trophy would not fit into the brown suitcase because it was too ⟨ ⟩. What was too ⟨ ⟩?

Answer 0: the trophy
Answer 1: the suitcase

special: big
alternate: small

The ⟨ ⟩ in a WS is a placeholder for the special word or its alternate, given in the first and second rows of the table below the line. A WS includes both the question and the answer: Answer 0 (the first party in the sentence) is the correct answer when the special word replaces the ⟨ ⟩ and Answer 1 (the second party) is the correct answer when the alternate word is used.

While a WS involves a pair of questions that have opposite answers, it is not necessary that the special word and its alternate be opposites (like "big" and "small"). Here are two examples where this is not the case:

- Paul tried to call George on the phone, but he was not ⟨ ⟩. Who was not ⟨ ⟩?

  Answer 0: Paul
  Answer 1: George

  special: successful
  alternate: available

- The lawyer asked the witness a question, but he was reluctant to ⟨ ⟩ it. Who was reluctant?

  Answer 0: the lawyer
  Answer 1: the witness

  special: repeat
  alternate: answer

In putting together an actual test for a subject, we would want to choose randomly between the special word and its

[2] See also the discussion of this in (Pylyshyn 1984).

alternate. Since each WS contains the two questions and their answers, a random WS test can be constructed, administered, and graded in a fully automated way. An expert judge is not required to interpret the results.

What is obvious?

The most problematic aspect of this proposed challenge is coming up with a list of appropriate questions. Like the RTE, candidate questions will need to be tested empirically before they are used in a test. We want normally-abled adults whose first language is English to find the answers obvious. But what do we mean by "obvious"? There are two specific pitfalls that we need to avoid.

Pitfall 1

The first pitfall concerns questions whose answers are in a certain sense too obvious. These are questions where the choice between the two parties can be made without considering the relationship between them expressed by the sentence. Consider the following WS:

The women stopped taking the pills because they were ⟨ ⟩. Which individuals were ⟨ ⟩?

Answer 0:	the women
Answer 1:	the pills
special:	pregnant
alternate:	carcinogenic

In this case, because only the women can be pregnant and only the pills can be carcinogenic, the questions can be answered by ignoring the sentence completely and merely finding the permissible links between the answers and the special word (or its alternate). In linguistics terminology, the anaphoric reference can be resolved using selectional restrictions alone. Because selectional restrictions like this might be learned by sampling a large enough corpus (that is, by confirming that the word "pregnant" occurs much more often close to "women" than close to "pills"), we should avoid this sort of question.

Along similar lines, consider the following WS:

The racecar zoomed by the school bus because it was going so ⟨ ⟩. What was going so ⟨ ⟩?

Answer 0:	the racecar
Answer 1:	the school bus
special:	fast
alternate:	slow

In principle, both a racecar and a school bus can be going fast. However, the association between racecars and speed is much stronger, and again this can provide a strong hint about the answer to the question. So it is much better to alter the example to something like the following:

The delivery truck zoomed by the school bus because it was going so ⟨ ⟩. What was going ⟨ ⟩?

Answer 0:	the delivery truck
Answer 1:	the school bus
special:	fast
alternate:	slow

As it turns out, this pitfall can also be avoided by only using examples with randomly chosen proper names of people (like Joan/Susan or Paul/George, above) where there is no chance of connecting one of the names to the special word or its alternate.

Pitfall 2

The second and more troubling pitfall concerns questions whose answers are not obvious enough. Informally, a good question for a WS is one that an untrained subject (your Aunt Edna, say) can answer immediately.

But to say that an answer is obvious does not mean that the other answer has to be *logically inconsistent*. It is possible that in a bizarre town, the councillors are advocating violence and choose to deny a permit as a way of expressing this. It is also possible that angry demonstrators could nonetheless fear violence and that the councillors could use this as a pretext to deny them a permit. But these interpretations are farfetched and will not trouble your Aunt Edna.[3] So they will not cause us statistical difficulties except perhaps with language experts asked to treat the example as an object of professional interest.

To see what can go wrong with a WS, however, let us consider an example that is a "near-miss." We start with the following:

Frank was jealous when Bill said that he was the winner of the competition. Who was the winner?

Answer 0:	Frank
Answer 1:	Bill

So far so good, with "jealous" as the special word and Bill as the clear winner. The difficulty is to find an alternate word that points to Frank as the obvious winner. Consider this:

Frank was pleased when Bill said that he was the winner of the competition.

The trouble here is that it is not unreasonable to imagine Frank being pleased because Bill won (and similarly for "happy" or "overjoyed"). The sentence is too ambiguous to be useful. If we insist on using a WS along these lines, here is a better version:

Frank felt ⟨ ⟩ when his longtime rival Bill revealed that he was the winner of the competition. Who was the winner?

Answer 0:	Frank
Answer 1:	Bill
special:	vindicated
alternate:	crushed

In this case, it is advisable to include the information that Bill was a longtime rival of Frank to make it more apparent that Frank was the winner.[4]

[3] Similarly, there is a farfetched reading where a small trophy would not "fit" in a big suitcase in the sense of fitting closely, the way a big shoe is not the right fit for a small foot.

[4] However, the vocabulary is perhaps too rich now.

Incremental Progress

In the end, what a subject will consider to be obvious will depend to a very large extent on what he or she knows. We can construct examples where very little needs to be known, like the trophy example, or this one:

> The man could not lift his son because he was so ⟨ ⟩. Who was ⟨ ⟩?
>
> | | |
> |---|---|
> | Answer 0: | the man |
> | Answer 1: | his son |
> | special: | weak |
> | alternate: | heavy |

At the other extreme, we have examples like the town councillor one proposed by Winograd. Unlike with the RTE, the "easier" questions are not easier because they can be answered in a more superficial way (using, for example, only statistical properties of the individual words). Rather, they differ on the background knowledge assumed. Consider, for example, this intermediate case:

> The large ball crashed right through the table because it was made of ⟨ ⟩. What was made of ⟨ ⟩?
>
> | | |
> |---|---|
> | Answer 0: | the ball |
> | Answer 1: | the table |
> | special: | steel |
> | alternate: | styrofoam |

For adults who know what styrofoam is, this WS is obvious. But for individuals who may have only heard the word a few times, there could be a problem.

It is perhaps an advantage of the WS challenge that like the RTE, it can be *staged*: we can have libraries of questions suitable for anyone who is at least ten-years old (like the trophy one), all the way up to questions that are more "university-level" (like the town councillor one). To get a feel for some of the possibilities, we include a number of additional examples in the Appendix at the end of the paper.

To help ensure that researchers can make progress on the WS challenge at first, we propose to make publicly available well beforehand a list of *all the words* that will appear in a test. (Of course, we would include both the special words and their alternates, although only one of them will be selected at random when the test is administered.) For a test with 20 questions, which should be more than enough to rule out mere guessing, 200 words (give or take proper names) should be sufficient. A test with 20 questions would only take a person 10 minutes or so to complete.

Discussion and Conclusion

The claim of this paper in its strongest form might be this: with a very high probability, anything that answers correctly a series of these questions (without having extracted any hints from the text of this paper, of course) is thinking in the full-bodied sense we usually reserve for people.

To defend this claim, however, we would have to defend a philosophical position that Turing sought to avoid with his original Turing Test. So like Turing, it is best to make a weaker claim: with a very high probability, anything that answers correctly is engaging in behaviour that we would say shows thinking in people. Whether or not a subject that passes the test is really and truly thinking is the philosophical question that Turing sidesteps.

It's not as if everyone agrees with Turing, however. Searle (1980) with his well-known Chinese Room thought experiment attempts to show that it is possible for people to get the observable behaviour right (in a way that would cover equally well the original Turing Test, an RTE test, and our WS challenge), but without having the associated mental attributes. However, in my opinion (Levesque 2009), his argument does not work properly.

On a related theme, Hawkins and Blakeslee (2004) suggest that AI has focussed too closely on getting the behaviour right and that this has prevented it from seeing the importance of what happens *internally* even when there is no external behaviour. The result, they argue, is a research programme that is much too behavioristic. (Searle makes a similar point.) See also (Cohen 2004).

In my opinion, this is a misreading of Turing and of AI research. Observable intelligent behaviour is indeed the ultimate goal according to Turing, but things do not stop there. The goal immediately raises a fundamental question: what sorts of computational mechanisms can possibly account for the production of this behaviour? And this question may well be answered in a principled and scientific way by postulating and testing for a variety of internal schemes and architectures. For example, what are we to make of a person who quietly reads a book with no external behaviour other than eye motion and turning pages? There can be a considerable gap between the time a piece of background knowledge is first acquired and the time it is actually needed to condition behaviour, such as producing the answer to a WS.

The computational architecture articulated by John McCarthy (1968) was perhaps the first to offer an even remotely plausible story about how to approach something like the WS challenge. This is what is sometimes called the *knowledge-based* approach (Brachman and Levesque 2004, Chap. 1). While the approach still faces tremendous scientific hurdles, it remains, arguably, the best game in town. However, nothing in the WS challenge insists on this approach; if statistics over a large corpus works better, so be it! What Turing sought to avoid is the philosophical discussion *assuming* we were able to produce the intelligent behaviour; but how we get there is wide open, including all sorts of internal activity when all is quiet on the external front.

So for our purposes, we can agree with Turing that getting the behaviour right is the primary concern. And we further agree that English comprehension in the broadest sense is an excellent indicator. Where we have a slight disagreement with Turing is whether a conversation in English is the right vehicle. Our WS challenge does not allow a subject to hide behind a smokescreen of verbal tricks, playfulness, or canned responses. Assuming a subject is willing to take a WS test at all, much will be learned quite unambiguously about the subject in a few minutes. What we have proposed here is certainly less demanding than an intelligent conversation about sonnets (say), as imagined by Turing; it does, however, offer a test challenge that is less subject to abuse.

References

D.G. Bobrow, C. Condoravdi, R. Crouch, V. de Paiva, L. Karttunen, T.H. King, R. Mairn, L. price, A. Zaenen, Precision-focussed textual inference in *Proc. of the Workshop on Textual Entailment and Paraphrasing* ACL, Prague, 2007.

R.J. Brachman and H.J. Levesque, *Knowledge Representation and Reasoning*. Morgan Kaufmann, 2004.

P.R. Cohen, If not the Turing Test, then what? Invited talk of AAAI-04, *xvi*, AAAI Press, 2004.

I. Dagan, O. Glickman, B. Magnini, The PASCAL recognising textual entailment challenge in *Machine Learning Challenges*, Springer Verlag, LNAI 3944, 2006.

S. Harnad, Minds, machines and Searle. *Journal of Theoretical and Experimental Artificial Intelligence* **1**, 1989.

J. Hawkins, S. Blakeslee, *On Intelligence*. Times Books, New York, 2004.

H.J. Levesque, Is it enough to get the behaviour right? *Proc. of IJCAI-09*, Pasadena, CA, 2009.

J. McCarthy, The advice taker. In M. Minsky, editor, *Semantic Information Processing*. MIT Press, 1968.

Z.W. Pylyshyn, *Computation and Cognition: Toward a Foundation for Cognitive Science*. MIT Press, 1984.

V. Rus, P. McCarthy, D. McNamara, A. Graesser, A study of textual entailment, *International Journal of Artificial Intelligence Tools*, **17**, 2007.

J. Searle, Minds, brains, and programs. *Brain and Behavioral Sciences* **3**, 417-457, 1980.

S.M. Shieber, Lessons from a restricted Turing Test. *CACM* **37**(6), 70–78, 1994.

A. Turing, Computing machinery and intelligence. *Mind* **59**, 433–460, 1950.

L. von Ahn, M. Blum, N. Hopper, J. Langford, CAPTCHA: Using Hard AI Problems for Security. *Advances in Cryptology, Eurocrypt 2003*, 294–311.

J. Weizenbaum, ELIZA. *CACM* **9**, 36–45, 1966.

T. Winograd, *Understanding Natural Language*. Academic Press, New York, 1972.

Appendix*

Here are additional Winograd schemas from which a question and its answer can be generated. See the text for details.

1. John could not see the stage with Billy in front of him because he is so ⟨⟩. Who is so ⟨⟩?

   Answer 0: John
   Answer 1: Billy

   special: short
   alternate: tall

2. Tom threw his schoolbag down to Ray after he reached the ⟨⟩ of the stairs. Who reached the ⟨⟩ of the stairs?

   Answer 0: Tom
   Answer 1: Ray

   special: top
   alternate: bottom

3. Although they ran at about the same speed, Sue beat Sally because she had such a ⟨⟩ start. Who had a ⟨⟩ start?

   Answer 0: Sue
   Answer 1: Sally

   special: good
   alternate: bad

4. The sculpture rolled off the shelf because it was not ⟨⟩. What was not ⟨⟩?

   Answer 0: the sculpture
   Answer 1: the shelf

   special: anchored
   alternate: level

5. Sam's drawing was hung just above Tina's and it did look much better with another one ⟨⟩ it. Which looked better?

   Answer 0: Sam's picture
   Answer 1: Tina's picture

   special: below
   alternate: above

6. Anna did a lot ⟨⟩ than her good friend Lucy on the test because she had studied so hard. Who studied hard?

   Answer 0: Anna
   Answer 1: Lucy

   special: better
   alternate: worse

7. The firemen arrived ⟨⟩ the police because they were coming from so far away. Who came from far away?

   Answer 0: the firemen
   Answer 1: the police

   special: after
   alternate: before

8. Frank was upset with Tom because the toaster he had ⟨⟩ him did not work. Who had ⟨⟩ the other party?

   Answer 0: Frank
   Answer 1: Tom

   special: bought from
   alternate: sold to

9. Jim ⟨⟩ Kevin because he was so upset. Who was upset?

   Answer 0: Jim
   Answer 1: Kevin

   special: yelled at
   alternate: comforted

10. The sack of potatoes had been placed ⟨⟩ the bag of flour, so it had to be moved first. What had to be moved first?

    Answer 0: a sack of potatoes
    Answer 1: a bag of flour

    special: on top of
    alternate: right under

*Thanks to Pat Levesque and reviewers for help with these examples and to Stavros Vassos for general discussion.