

Tense Happens to be a Good Predictor for Implicative Verbs: Supplementary Material

Ellie Pavlick

University of Pennsylvania
epavlick@seas.upenn.edu

Chris Callison-Burch

University of Pennsylvania
ccb@cis.upenn.edu

1 Heuristic algorithm for determining tense of the main clause

Algorithm 1 Procedure for assigning tense to main verb, v at index i .

```
pos ← fine-grained POS tag of v
m1, m2 ← words at indices  $i - 1$  and  $i - 2$ 
if m1 is a negation (e.g. not) then
    m1 ← m2           ▷ Skip over negations
end if
if m1 in past then:
    tense = past
else if m1 in modals or future or m2 in future
then:
    tense = future
else if m2 in modals then
    if stem of m1 == have then
        tense = past
    else:
        tense = future
    end if
else if pos == VB or VBG or VBP or VBZ then
    tense = present
else if pos == VBD then
    tense = past
else if pos == VBN then
    if m1 == is or are then
        tense = present
    else if m1 == was or were then
        tense = past
    else
        tense = past
    end if
end if
```

The above algorithm shows the heuristic procedure used to assign a tense to the main verb of a sentence. This procedure relies on the word lists shown in Table 1. Ideally, the tense-agreement method described in this paper could be made more robust if this heuristic procedure were replaced with a statistical model for learning coarse-grained (past/present/future) verb tense.

future	<i>will, wo</i>
past	<i>did, do, had, were</i>
modals	<i>can, could, may, might, must, shall, should, wo, would</i>
negations	<i>n't, not</i>

Table 1: The following word lists are used in Algorithm 1 below in order to determine the tense of the main clause. Note some odd words (e.g. *wo*) arise from stemming of contractions.

2 Annotation Guidelines

We collect entailment annotations on a 5-point scale, ranging from 1 (definite contradiction) to 5 (definite entailment), with 2 and 4 capturing likely (but not certain) contradiction/entailment respectively. We recruit annotators on Amazon Mechanical Turk. We tell each annotator to assume that the premise is true, or describes a real scenario and then, using their best judgement, to indicate how likely it is, on a scale of 1 to 5, that the hypothesis is also true, or describes the same scenario. Workers are given the option to say that the sentence does not make sense, to account for poorly constructed p/h pairs, or errors in our parsing. A screenshot is shown in Figure 1 and the full annotation guidelines are given in Table 2.

30%

In fact, Americans, in his mind, **prefer to focus on** the sideshows and ignore the big picture.

In fact, Americans, in his mind, **focus on** the sideshows and ignore the big picture.

✓ --

- 5: The sentence is definitely true.
- 4: The sentence is probably true. It is more like to be true than false given the information available.
- 3: The sentence is not necessarily true or necessarily false. It is not possible to say given the information available.
- 2: The sentence is probably NOT true. It is more like to be false than true given the information available.
- 1: The sentence is definitely NOT true

At least one of the sentences does not make sense.

Figure 1: Screenshot of annotation interface shown to workers.

We ultimately collapse the 5-point scores into a three class annotation, in which ≤ 2 indicates *contradiction*, ≥ 4 indicates *entailment*, and 3 indicates *neutral/unknown*. Inter-annotator agreement is $\kappa=0.52$ for this 3-way annotation. The full results of the annotation for all 77 verbs which were labeled are shown in Table 3.

We are researchers from the Anonymous, trying to help computers understand language! Understanding language allows computers to do useful things like answer our questions or summarize news articles for us. While common sense reasoning is easy for people, it is very very hard for computers. Please help us by using your common sense to decide if sentences are more likely to be true or false.

Note: The sentences have been automatically generated. Please try to **interpret them to the best of your ability**, even though they may contain grammatical errors and typos. **You can always select the "does not make sense" option** if you feel you cannot make any reasonable inference of what the sentence is supposed to mean. You will never be punished for choosing this option.

Instructions For each pair of sentences, assume that the first sentence is true, describes a real scenario, or expresses an opinion. Using your best judgement, indicate how likely it is, on a scale of 1 to 5, that the second sentence is also true, describes the same scenario, or expresses the same opinion.

1. Your answers should be based only on information which is stated or implied by the first sentence. *Even if the second sentence seems like it is reasonably true in general, you should only choose 4 or 5 if the truth of the second sentence can be inferred entirely from the first sentence. E.g. for the sentence pair below, you should choose 3 since the first sentence alone provides us no way of knowing that Greece is in Europe (even though our world knowledge tells us that the second sentence is true).*

Example 1

Greece is a country.

Greece is a European country.

CORRECT ANSWER: 3, not necessarily true or necessarily false.

2. The order of the sentences is important. *For example, when the same sentences as above are provided in the opposite order, the correct answer changes to 5: definitely true.*

Example 2

Greece is a European country.

Greece is a country.

CORRECT ANSWER: 5, definitely true.

3. It is okay to make reasonable assumptions. *If the first sentence makes it highly unlikely that the second is true, indicate so by choosing option 1 or 2. Try to interpret the sentences as you would if you heard them in a real life conversation. It is okay to make reasonable assumptions that you believe most people would make. E.g. it is okay to assume that a bomb is not a fake bomb in the below context.*

Example 3

The terrorists were collecting materials to build a bomb.

The terrorists were collecting materials to build a fake bomb.

CORRECT ANSWER: 1, definitely NOT true. or 2, probably NOT true.

Example 4

The man failed to submit the report on time

The man submitted the report on time

CORRECT ANSWER: 1, definitely NOT true. or 2, probably NOT true.

4. When in doubt, you should err on the side of uncertainty. *If the context of the first sentence makes it reasonably possible for the second to be either true or false, choose 3.*

Example 5

Secretary Clinton is the expected nominee.

Secretary Clinton is the nominee.

CORRECT ANSWER: 3, not necessarily true or necessarily false.

Example 6

The police have arrested a suspected murderer.

The police have arrested a murderer.

CORRECT ANSWER: 3, not necessarily true or necessarily false.

5. Take the entire sentence into account. *Remember that we are interested in the information communicated by the sentence as a whole. Inserting the same words might lead to different answers in different contexts.*

Example 7

She is a potential candidate for the Senate.

She is a candidate for the Senate.

CORRECT ANSWER: 3, not necessarily true or necessarily false.

Example 8

They are talking about potential candidates for the Senate.

They are talking about candidates for the Senate.

CORRECT ANSWER: 4, probably true or 5, definitely true.

Keep in mind, we are **predominantly interested in understanding whether the second sentence communicates the same information as the first, or if it adds or removes important information.**

Table 2: Full annotation guidelines shown to annotators on Amazon Mechanical Turk. Before gaining access to our annotation tasks, workers were required to achieve perfect accuracy on a qualification test consisting of 12 questions very similar to the examples shown in these instructions.

0.8782	force to	E	E:***** N:* C:**	0.5455	choose to	E	E:***** N:** C:	0.2363	plot to	N	E:*** N:***** C:**
0.8676	send to	E*	E:**** N:**** C:*	0.5363	return to	E	E:***** N: C:*	0.2353	favor to	E	E:***** N:**** C:*
0.8333	appoint to	C	E:** N:* C:*****	0.5256	believe to	E	E:***** N:***** C:***	0.2329	push to	N	E:***** N:***** C:**
0.8326	begin to	E	E:***** N:* C:	0.5248	call to	E	E:***** N:**** C:	0.2241	battle to	N	E:***** N:***** C:*
0.8120	admit to	E	E:***** N:* C:***	0.5246	hold to	E	E:***** N:* C:	0.2235	meet to	E	E:***** N: C:
0.7984	confess to	E	E:***** N:***** C:	0.5229	get to	E	E:***** N:* C:	0.2023	work to	E	E:***** N:**** C:
0.7949	start to	E	E:***** N:*** C:	0.4940	opt to	E	E:***** N:* C:	0.1991	vow to	N	E:***** N:***** C:
0.7900	manage to	E	E:***** N: C:	0.4921	help to	E	E:***** N:** C:**	0.1832	bid to	N	E:*** N:***** C:***
0.7798	apply to	C	E:*** N: C:*****	0.4586	attempt to	N	E:**** N:***** C:**	0.1719	fight to	E	E:***** N:***** C:
0.7512	like to	E	E:***** N: C:	0.4464	decide to	E	E:***** N:** C:	0.1620	determine to	E	E:***** N:* C:
0.7470	elect to	E	E:***** N:* C:	0.4228	mean to	E	E:***** N:***** C:	0.1596	wait to	C	E:*** N:*** C:*****
0.7414	continue to	E	E:***** N: C:	0.4184	suppose to	C	E:**** N:**** C:***	0.1579	pledge to	N	E:*** N:***** C:*
0.7344	decline to	C	E:* N: C:*****	0.4169	try to	N	E:*** N:***** C:	0.1394	commit to	E	E:***** N:***** C:
0.7222	seem to	E	E:***** N:**** C:	0.4094	poise to	N	E:***** N:***** C:	0.1360	propose to	N	E:*** N:***** C:***
0.7198	come to	E	E:***** N:** C:	0.3592	threaten to	N	E:***** N:***** C:*	0.1331	expect to	N	E:*** N:***** C:**
0.7153	appear to	E	E:***** N:***** C:	0.3571	register to	N	E:** N:***** C:****	0.1277	set to	N*	E:* N:*** C:*
0.6835	move to	N	E:***** N:***** C:	0.3548	order to	E	E:***** N:** C:*	0.1235	forecast to	N	E:** N:***** C:
0.6833	allow to	E	E:***** N:* C:	0.3478	tell to	N*	E:** N:***** C:*	0.1186	wish to	E	E:***** N:*** C:
0.6705	permit to	N*	E:*** N:***** C:**	0.3438	agree to	E	E:***** N: C:	0.1151	intend to	N	E:***** N:***** C:*
0.6574	refuse to	C	E: N:* C:*****	0.3350	require to	N	E:**** N:***** C:**	0.1051	prepare to	N	E:***** N:***** C:
0.6435	leave to	E	E:***** N: C:*	0.3333	prefer to	E	E:***** N:*** C:	0.1007	look to	E	E:***** N:**** C:
0.6222	offer to	E	E:***** N:***** C:	0.3333	rush to	E	E:***** N:*** C:*	0.0961	plan to	N	E:*** N:***** C:*
0.6130	vote to	E	E:***** N:***** C:	0.3100	seek to	E	E:***** N:**** C:**	0.0468	aim to	N	E:***** N:***** C:*
0.5837	fail to	C	E:*** N: C:*****	0.3018	struggle to	E	E:***** N:**** C:*****	0.0437	project to	N*	E: N:**** C:
0.5506	make to	E*	E:***** N:** C:**	0.2871	need to	E	E:***** N:**** C:***	0.0272	hope to	N	E:*** N:***** C:*
0.5500	ask to	E	E:***** N:**** C:***	0.2597	gather to	E	E:***** N: C:*				

Table 3: The full list of verbs which were labeled on MTurk. Columns in order are: tense agreement score, the verb, the majority entailment judgement, the distribution over entailment judgements. Each verb was judged in 3 contexts and each context received 5 independent judgements, for a total of 15 judgements per verb. Judgements were coarsely grouped into contradiction (C), neutral (N), or entailment (E). Annotators also had the option of choosing “NA” if they felt the premise/hypothesis pair did not make sufficient sense to be judged. For space reasons, “NA” judgements are not shown as part of the distribution. Those marked with * in the third column actually received a majority judgement of NA, but the second most frequent class is shown.