Machine Translation-style Textual Entailment

Sebastian Pado

October 20, 2012

Disclaimer: the original idea of this approach is due to Gil.

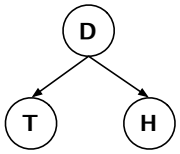
1 Preliminaries

- A probabilistic approach to TE
- Given: a set of documents describing "situations in the world"
- Goal: define $P(H)$, $P(T)$, $P(H|T)$ in terms of these documents
- Then define "faith" in the entailment as either $P(H|T) - P(H)$ ("additive faith") or $P(H|T)/P(H)$ ("multiplicative faith") – both quantities reflect the increase in the likelihood of seeing H that is caused by having seen T.
  - NOTE - I just realized that the "multiplicative faith" is the same concept as the pointwise mutual information between T and H (modulo logarithmization). Let's keep this in mind.
- My intuition: this a fairly straightforward application of some generative process that related D, T, and H.
- Structure of this draft: define a set of models that break down the process in terms of priors and conditional probabilities (Bayes Net-style). Compare these models on an abstract level, without caring about how to actually instantiate the conditional probabilities.
- Then discuss the instantiation

2 Model 1

- Intuition: Treat documents as descriptions of situations.
- Generative story:
  - Documents are (not-so-)latent classes
  - All documents are equally likely ($1/|D|$)

– Documents generate both texts and hypotheses



– This means that longer texts generate shorter text: it's easy to imagine a straightforward sampling/selection process underlying the generation

– The joint PDF is straightforward:

$$P_1(H, D, T) = P(D)P(H|D)P(T|D) \tag{1}$$

Because we consider *all* documents, we marginalize out over D:

$$P_1(H, T) = \sum_i P(D_i)P(H|D_i)P(T|D_i) = \frac{1}{|D|}\sum_i P(H|D_i)P(T|D_i) \tag{2}$$

– That's all.

- The first quantity of interest is $P(H)$, the probability of H given the document collection. Our generative story defines it as

$$P_1(H) = \sum_i P(D)P(H|D_i) = \frac{1}{|D|}\sum_i P(H|D_i) \tag{3}$$

- The second quantity of interest is $P(H|D, T)$, the probability of H given T and the document collection. Our generative story defines it as

$$P_1(H|T) = \frac{P(H,T)}{P(T)} = \frac{\sum_i P(H,T|D_i)}{\sum_i P(T|D_i)} = \frac{|D|}{|D|}\frac{\sum_i P(H|D_i)P(T|D_i)}{\sum_i P(T|D_i)} \tag{4}$$

$$= \frac{\sum_i P(H|D_i)P(T|D_i)}{\sum_i P(T|D_i)} \tag{5}$$

- The additive faith $P_1(H|T) - P_1(H)$ then comes out as:

$$P_1(H|T) - P_1(H) = \frac{\sum_i P_1(H|D_i)P_1(T|D_i)}{\sum_i P_1(T|D_i)} - \frac{1}{|D|}\sum_i P_1(H|D_i) \tag{6}$$

This expression can be understood as follows: Both $P(H|T)$ and $P(H)$ are averages over $P(H|D_i)$, that is, quantify "how probable" H is on average in the documents. The difference is that $P(H|T)$ *weighs* each documents by its $P(T|D)$, that is, by its probability of generating T, while $P(H)$ assumes that the weight of each document is 1. (The two terms become identical if $P(T|D)$ is 1 for all D). That means that $P(H|T) - P(H)$ is positive if $P(H|D)$ is larger for those documents where $P(T|D)$ is larger.

- The multiplicative faith $P_1(H|T)/P_1(H)$ is:

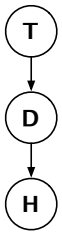
$$P_1(H|T)/P_1(H) = |D|\frac{\sum_i[P_1(H|D_i)P_1(T|D_i)]}{\sum_i[P_1(T|D_i)]\sum_i[P_1(H|D_i)]} \tag{7}$$

  As I said above, this is the (exponentiated) PMI between H and T.

- There is one problem with this model though: It assumes exchangeability of H and T. So $P(H|D,T) = P(T|D,H)$. This basically means that it is symmetrical, which may be appropriate for a paraphrase model, but definitely not for an entailment model.

3 Model 2

- So let's try to formulate a model that does not share this problem.
- New generative story:
  - We start with a text
  - The text generates documents
  - The documents generate the hypothesis



- Joint PDF:

$$P_2(H,T) = P_2(T)\sum_i[P_2(D_i|T)P_2(H|D_i)] \tag{8}$$

- What seems weird that we have once a short text generating a long text and once a long text generating a short text. Let's see if we can reformulate the expression using Bayes' rule.

$$P_2(H,T) = P_2(T)\sum_i[\frac{P_2(T|D_i)P_2(D_i)}{P_2(T)}P_2(H|D_i)] \tag{9}$$

$$= \sum_i[P_2(T|D_i)P_2(D_i)P_2(H|D_i)] \tag{10}$$

This formula contains a prior for the documents. We will assume again that it is uniform. Okay; what do $P_2(H|T)$ and $P_2(H)$ look like now?

$$P_2(H|T) = \frac{P_2(H,T)}{P_2(T)} = \frac{1}{|D|P(T)} \sum_i [P_2(T|D_i)P_2(H|D_i)] \tag{11}$$

$$P_2(H) = \sum_i P_2(D_i)P_2(H|D_i) = \frac{1}{|D|} \sum_i P_2(H|D_i) \tag{12}$$

Our additive faith in the entailment is

$$P_2(H|T) - P_2(H) = \frac{1}{|D|} \sum_i [P_2(H|D_i) \cdot (\frac{P_2(T|D_i)}{P(T)} - 1)] \tag{13}$$

Let's try to analyse this. The formula says that our faith in H given T is high if there are documents which are likely to generate H and whose probability of generating T is bigger than the prior probability of T.

  - What's interesting is the comparison to the same quantity in model 1 (Eq. 6): The two formulae are basically parallel. They differ in *when* normalization across documents takes place: In Model 1, the global evidence is normalized; in Model 2, we normalize the evidence within each document. I need to think more about what that means.

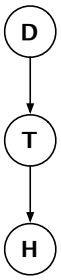
- The multiplicative faith is:

$$P_2(H|T)/P_2(H) = \frac{1}{P(T)} \frac{\sum_i P_2(T|D_i)P_2(H|D_i)}{\sum_i P_2(H|D_i)} \tag{14}$$

That's fairly straightforward to interpret: We divide the probability that the documents generate both T and H by the probability that they generate just H.

4 Model 3

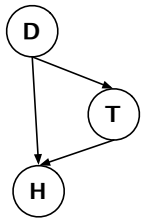
- Generative story:
  - Start from the document
  - Document generates the text
  - The text generates the hypothesis



Problem: Due to the conditional independence, $P(H|T)$ is not influenced by D. So we can forget about this model.

5 Model 4

- Generative story:
  - Start from the document
  - Document generates Text
  - Text and document generate Hypothesis together



- Joint PDF:

$$P(H,T) = \sum_i P(D_i)P(T|D_i)P(H|D_i,T) \tag{15}$$

$$\tag{16}$$

With a uniform prior for documents:

$$P(H,T) = \frac{1}{|D|}\sum_i P(T|D_i)P(H|D_i,T) \tag{17}$$

$$\tag{18}$$

So what we're getting in this case is a model that averages over all documents, keeping track of how likely the document is to generate T, and how likely the document plus T is to generate H. The last term can be understood in terms of "query expansion", that is, we allow D_i and T to combine forces in generating H.

- Important quantities:

$$P(H|T) = P(H,T)/P(T) = \frac{\sum_i P(T|D_i)P(H|D_i,T)}{\sum_i P(T|D_i)} \tag{19}$$

$$P(H) = \sum_{i,j} P(H|D_i,T_j) \tag{20}$$

The problem in this model is $P(H)$ because to estimate it we need to sum over all pairs of documents and texts (which we don't know).

- Entailment Faith: the formulae are easy to write down with the above, but are difficult to interpret because $P(H)$ includes the double summation.

6 Query Rewriting

- The seminal paper by Berger and Lafferty (1999) defines a family of models exactly for the task that we face in models 1 and 2 (create a sentence given a document). They are very simple, though. – NB - they rely to a large extent on language models ;-).
- Riezler and Liu (2010) present a more recent variation on this theme which involves more recent machine translation techniques to translate "from the language of user queries to the language of web documents".
- What these methods require, though, are parallel corpora. They get them from query logs – we would have to pair Ts and Hs with web documents. It's not clear to me where we'd get such data from, because that's exactly the task that entailment is suppose to solve (Well, we might just re-use existing models of this type if they are available somewhere.)
- What's notable is that in none of the models we really get $P(H|T)$ because apparently that's hard to combine sensibly with the document idea – i.e. if we can estimate $P(H|T)$ directly, how would knowledge about documents influence that?