A Probabilistic Lexical Approach to Textual Entailment*

Oren Glickman, Ido Dagan, Moshe Koppel
Bar Ilan University
Computer Science Department
Ramat-Gan, Iarael
{glikmao, dagan, koppel}@cs.biu.ac.il

Abstract

The textual entailment problem is to determine if a given text entails a given hypothesis. This paper describes first a general generative probabilistic setting for textual entailment. We then focus on the sub-task of recognizing whether the lexical concepts present in the hypothesis are entailed from the text. This problem is recast as one of text categorization in which the classes are the vocabulary words. We make novel use of Naïve Bayes to model the problem in an entirely unsupervised fashion. Empirical tests suggest that the method is effective and compares favorably with state-of-the-art heuristic scoring approaches.

Textual Entailment

Many Natural Language Processing (NLP) applications need to recognize when the meaning of one text can be expressed by, or inferred from, another text. Information Retrieval (IR), Question Answering (QA), Information Extraction (IE) and text summarization are examples of applications that need to assess such semantic overlap between text segments. Textual Entailment Recognition has recently been proposed as an application independent task to capture such semantic inferences and variability [Dagan et al., 2005]. A text *t* textually entails a hypothesis *h* if *t* implies the truth of *h*. Textual entailment captures generically a broad range of inferences that are relevant for multiple applications. For example, a QA system has to identify texts that entail the expected answer. Given the question "Where was Harry Reasoner born?", a text that includes the sentence "*Harry Reasoner's birthplace is Iowa*" entails the expected answer form "Harry Reasoner was born in Iowa." In many cases, though, entailment inference is uncertain and has a probabilistic nature. For example, a text that includes the sentence "*Harry Reasoner is returning to his Iowa hometown to get married.*" does not deterministically entail the above answer form. Yet, it is clear that it does add substantial information about the correctness of the hypothesized assertion.

A Probabilistic Setting

We propose a general generative probabilistic setting for textual entailment. We assume that a language source generates texts within the context of some state of affairs. Thus, texts are generated along with hidden truth assignments to hypotheses. We define two types of events over the corresponding probability space:

I) For a hypothesis *h*, we denote as Tr_h the random variable whose value is the truth value assigned to *h* in the world of the generated text. Correspondingly, $\mathrm{Tr}_h=1$ is the event of *h* being assigned a truth value of 1 (True).

II) For a text *t*, we use *t* to denote also the event that the generated text is *t*.

Textual entailment relationship: We say that *t* probabilistically entails *h* (denoted as $t \Rightarrow h$) if *t* increases the likelihood of *h* being true, that is if $P(\mathrm{Tr}_h = 1| t) > P(\mathrm{Tr}_h = 1)$.

Entailment confidence: We quantify the marginal amount of information contributed by the text to assessing the truth of the hypothesis relative to its prior with the pointwise mutual information: $I(\mathrm{Tr}_h=1,t)=\log(P(\mathrm{Tr}_h = 1| t) / P(\mathrm{Tr}_h = 1))$.

An Unsupervised Lexical Model

The proposed setting above provides the necessary grounding for probabilistic modeling of textual entailment. However, it is important to bear in mind that it is not trivial to estimate the constituent probabilities in the definition of textual entailment since the truth assignments of hypotheses for the corpus' texts are not observed.

Lexical Entailment as Text classification

As modeling the full extent of the textual entailment problem is a long term research goal, we focus on a sub task we term lexical entailment - recognizing if the individual lexical concepts in a hypothesis are entailed from a given text.

When estimating the entailment probability we assume that the truth probability of a term in a hypothesis *h* is independent of the truth of the other terms in *h*, obtaining:

$$P(\mathrm{Tr}_h = 1| t) = \Pi_{u\in h}P(\mathrm{Tr}_u=1|t)$$
$$P(\mathrm{Tr}_h = 1) = \Pi_{u\in h}P(\mathrm{Tr}_u=1) \qquad (1)$$

At this point, it is perhaps best to think of the entailment problem as a text classification task in which the classes are an abstract binary notion of lexical truth (for the different words in the vocabulary). First, we construct the initial la-

* *This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.*

beling based solely on the explicit presence or absence of each u in t. Then we apply Naïve Bayes in an unsupervised fashion that derives analytically from the defined probabilistic setting.

Initial Labeling

As an initial approximation, we assume that for any document in the corpus the truth value corresponding to a term u is determined by the explicit presence or absence of u in that document.

In some respects the initial labeling is similar to systems that perform a Boolean search (with no expansion) on the keywords of a textual hypothesis in order to find candidate (entailing) texts. Of course, due to the semantic variability of language, similar meanings could be expressed in different wordings, which is addressed in the subsequent model. The initial labeling, however, may provide useful estimates for this model.

Naïve Bayes Refinement

Following the standard naïve Bayes assumption, we can rewrite the probability $\mathrm{P}(\mathrm{Tr}_u\!=\!1|t)$ as in (2). In this way we are able to estimate $\mathrm{P}(\mathrm{Tr}_u\!=\!1|t)$ based on the prior $\mathrm{P}(v|\ \mathrm{Tr}_u\!=\!1)$ and the lexical probabilities $\mathrm{P}(v|\ \mathrm{Tr}_u\!=\!1)$ and $\mathrm{P}(v|\ \mathrm{Tr}_u\!=\!0)$ for every u, v in the vocabulary V. These probabilities are easily estimated from the corpus given the initial model's estimate of truth assignments, assuming a multinomial event model for documents and Laplace smoothing ([McCallum and Nigam, 1998]).

$$\mathrm{P}(\mathrm{Tr}_u = 1 \mid t) = \frac{\mathrm{P}(\mathrm{Tr}_u = 1)\prod_{v\in t}\mathrm{P}(v \mid \mathrm{Tr}_u = 1)}{\sum_{c\in\{0,1\}}\mathrm{P}(\mathrm{Tr}_u = c)\prod_{v\in t}\mathrm{P}(v \mid \mathrm{Tr}_u = c)} \qquad (2)$$

From above equations we have a refined probability estimate for $\mathrm{P}(\mathrm{Tr}_h\!=\!1|\ t)$ and $\mathrm{P}(\mathrm{Tr}_h\!=\!1)$ for any arbitrary text t and hypothesis h. The criterion for turning probability estimates into classification decisions is derived analytically from our proposed probabilistic setting of textual entailment. We classify positively for entailment if $\mathrm{P}(\mathrm{Tr}_h\!=\!1|\ t) > \mathrm{P}(\mathrm{Tr}_h\!=\!1)$ and assign a confidence score of $\log(\mathrm{P}(\mathrm{Tr}_h\!=\!1|t)\ /\ \mathrm{P}(\mathrm{Tr}_h\!=\!1))$ for ranking purposes. In fact, the empirical evaluation showed this analytic threshold to be almost optimal.

1 Empirical Evaluation

Experimental Setup

Though empirical modeling of semantic inferences between texts is commonly done within application settings, there is no common dataset available to specifically evaluate a textual entailment system. In order to test our model we therefore needed an appropriate set of text-hypothesis pairs. We chose the information seeking setting, common in applications such as QA and IR, in which a hypothesis is given and it is necessary to identify texts that entail it. The evaluation criterion is application-independent based on human judgment of textual entailment.

Experiments were done on the *Reuters Corpus Volume 1*. An annotator chose 50 sentential hypotheses from the corpus sentences. We required that the hypotheses convey a reasonable information need in such a way that they might correspond to potential questions, semantic queries or IE relations. We created a set of candidate entailing texts for the given set of test hypotheses, by following common practice of morphological and WordNet-based expansion. Boolean search (with expanded words) was performed at the paragraph level over the full Reuters corpus. Paragraphs containing all original words of the hypothesis or their morphological derivations were excluded from the result set and selected a random set of 20 texts for each of the hypotheses.

The resulting dataset was given to two judges to be annotated for entailment. Corresponding to the notion of textual entailment, judges were asked to annotate a text-hypothesis pair as true if, given the text, they could infer with some confidence that the hypothesis is true. They were instructed to annotate the example as false if either they believed the hypothesis to be false given the text or if the text is unrelated to the hypothesis. A subset of 200 pairs was cross-annotated for agreement, resulting in a moderate Kappa statistic of 0.6. Overall, the annotators deemed 48% of the text-hypothesis pairs as positive examples of entailment.

Empirical Results

We trained our model on the Reuters corpus, classified the text-hypothesis pairs of the dataset and compared the model's prediction with the human judgments. The resulting (macro) average accuracy was 70%. Since our dataset did not include texts containing all content words of the hypotheses, the baseline model would have predicted none of the pairs to be correct (i.e. the text entails the hypothesis) yielding an average accuracy of only 52%.

We also compared our system's ranking ability. The entailment confidence score was used to rank the various texts of each hypothesis. The average *confidence weighted score* (*cws*) was measured for each hypothesis. The resulting cws macro average was 0.54 compared to average cws of 0.49 for random ordering. For further comparison, a state of the art *idf* semantic overlap measure [Saggion et al., 2004; Monz and de Rijke, 2001] achieved a score of 0.51. Though our model performs just slightly better, the results are statistically significant at the 0.02 level.

References

[Dagan et al., 2005] Ido Dagan, Oren Glickman and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. *PASCAL Challenges Workshop*.

[McCallum and Nigam, 1998] Andrew McCallum and Kamal Nigam. A Comparison of Event Models for Naive Bayes Text Classification. *AAAI-98 Workshop on "Learning for Text Categorization"*.

[Munz and de Rijke, 2001] Christof Monz, Maarten de Rijke. Light-Weight Entailment Checking for Computational Semantics. *The third workshop on inference in computational semantics (ICoS-3)*.

[Saggion et al., 2004] Horacio Saggion, Rob Gaizauskas, Mark Hepple, Ian Roberts and Mark A. Greenwood. Exploring the Performance of Boolean Retrieval Strategies For Open Domain Question Answering. *SIGIR04 Workshop on Information Retrieval for Question Answering*.