Web Based Probabilistic Textual Entailment

Oren Glickman, Ido Dagan and Moshe Koppel

Computer Science Department
Bar Ilan University
Ramat Gan, Israel

{glikmao,dagan,koppel}@cs.biu.ac.il

Abstract

This paper proposes a general probabilistic setting that formalizes the notion of textual entailment. In addition we describe a concrete model for lexical entailment based on web co-occurrence statistics in a bag of words representation.

1 Introduction

This paper describes the Bar-Ilan system participating in the Recognising Textual Entailment Challenge[1]. We first propose a general probabilistic setting that formalizes the notion of textual entailment. We then describe a model, derived from the proposed probabilistic setting, for lexical entailment based on web co-occurrence statistics in a bag of words representation.

Although our proposed lexical system is relatively simple, as it doesn't rely on syntactic or other deeper analysis, it nevertheless achieved an overall accuracy of 59% and an average precision of 0.57. The system did particularly well on Comparable Documents (CD) task achieving an accuracy of 83%. The results suggest that the proposed probabilistic framework is a promising basis for improved implementations that incorporate richer information.

2 Probabilistic Textual Entailment

2.1 Motivation

In many intuitive cases, the textual entailment recognition task may be perceived as being deterministic (Dagan and Glickman, 2004). For example, given the hypothesis h_1 = "*Harry was born in Iowa*" and a candidate text t_1 that includes the sentence "*Harry's birthplace is Iowa*", it is clear that t_1 does (deterministically) entail h_1, and humans are likely to have high agreement regarding this decision. In many other texts, though, entailment inference is uncertain and has a probabilistic nature. For example, a text t_2 that includes the sentence "*Harry is returning to his Iowa hometown to get married.*" does not deterministically entail the above h_1 since Harry might have moved to Iowa as a child. Yet, it is clear that t_2 does add substantial information about the correctness of h_1. In other words, the probability that h_1 is indeed true given the text t_2 ought to be significantly higher than the prior probability of h_1 being true. More specifically, we might say that the probability p of h_1 being true should be estimated based on the percentage of cases in which someone's reported hometown is indeed his/her birthplace. Accordingly, we wouldn't accept t_2 as a definite assessment for the truth of h_1. However, in the absence of other definite information, t_2 may partly satisfy our information need for an assessment of the probable truth of h_1, with p providing a confidence probability for this inference.

In the next section we propose a concrete probabilistic setting that formalizes the above rational.

2.1 A Probabilistic Setting

Let T denote a space of possible texts, and $t \in T$ a specific text.

Meanings are captured in our model by hypotheses and their truth values. Let H denote the set of all possible *hypotheses*. A hypothesis $h \in H$ is a propositional statement which can be assigned a truth value. For now it is assumed that h is represented as a textual statement, but in principle other representations for h may fit our framework as well. (For example, h might be syntactically/semantically annotated and possibly include Prolog-style existentially quantified variables).

---

[1] http://www.pascal-network.org/Challenges/RTE/

A semantic state of affairs is captured by a *possible world w: $H \rightarrow \{0, 1\}$*, which is defined as a mapping from H to {0=False, 1=True}, representing the set of w's concrete truth value assignments for all possible propositions. Accordingly, W denotes the set of all possible worlds.

A Generative Model

We assume a probabilistic generative model for texts and possible worlds. In particular, we assume that texts are generated within the context of some state of affairs, represented by a possible world. Thus, whenever the source generates a text t, it generates also hidden truth assignments that constitute a possible world w. The hidden w is perceived as a "snapshot" of the (complete) state of affairs in the world within which t was generated.

The probability distribution of the source, over all possible texts and truth assignments $T \times W$, is assumed to reflect only inferences that are based on the generated texts. That is, we assume that the distribution of truth assignments is not bound to reflect the state of affairs in any "real" world, but only the inferences about propositions' truth that are related to the text. In particular, the probability for generating a true hypothesis h that is not related at all to the corresponding text is determined by some prior probability $P(h)$, which is not bound to reflect h's prior in the "real" world. For example, h="Paris is the capital of France" might have a prior smaller than 1 and might well be false when the generated text is not related at all to Paris. In fact, we may as well assume that $P(h) = 1$ only for logical tautologies. On the other hand, we assume that the probability of h being true (generated within w) would be higher than the prior when the corresponding t does contribute information that supports h's truth.

We define two types of events over the probability space for $T \times W$:
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, $\text{Tr}_h=1$ 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 (as usual, it is clear from the context whether t denotes the text or the corresponding event).

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 $P(\text{Tr}_h = 1 | t) > P(\text{Tr}_h = 1)$ -- or equivalently if the pointwise mutual information, $I(\text{Tr}_h=1,t)$, is greater than 1.

Entailment confidence

Once *knowing* that $t \Rightarrow h$, we are further interested in a probabilistic confidence value for h being true given t, which corresponds to $P(\text{Tr}_h = 1 | t)$.

3   Lexical Entailment Models

The proposed setting above provides the necessary grounding for probabilistic modeling of textual entailment. As modeling the full extent of the textual entailment problem is a long term research goal, we focus here on identifying when the lexical elements of a textual hypothesis h are inferred from a given text t, even if the relations between these concepts may not be entailed from t.

To model lexical entailment we first assume that the meanings of the individual (content) words in a hypothesis $h=\{u_1, \ldots, u_m\}$ can be assigned truth values. A possible interpretation for these truth values, common in formal semantics tradition, is that lexical concepts are assigned existential meanings. For example, for a given text t, $\text{Tr}_{acquired}=1$ if it can be inferred in t's state of affairs that an acquisition event exists (occurred). It is important to note though that this is one possible interpretation. We only assume that truth values are defined for lexical items, but do not explicitly annotate or evaluate this sub-task.

Given this setting, a hypothesis is assumed to be true if and only if all its lexical components are true. 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:

$$P(Tr_h = 1 | t) = \prod_{i=1}^{m} P(Tr_{u_i} = 1 | t) \quad (1)$$

In order to estimate $P(\text{Tr}_u=1|v_1, \ldots, v_n)$ for a given word u and text $t=\{v_1, \ldots, v_n\}$, we further assume that the majority of the probability mass comes from a specific entailing word in t:

$$P(Tr_u = 1 | t) = \max_{v \in t} P(Tr_u = 1 | T_v) \quad (2)$$

where T_v denotes the event that a generated text contains the word v. This corresponds to expecting that each word in h will be entailed from a specific

word in t (rather than from the accumulative context of t as a whole). Alternatively, one can view (2) as inducing an alignment between the terms in h to the terms in the t, somewhat similar to alignment models in statistical MT (Brown et al., 1993).

Thus we propose estimating the entailment probability based on lexical entailment probabilities from (1) and (2) as follows:

$$P(Tr_h = 1 | t) = \prod_{u \in h} \max_{v \in t} P(Tr_u = 1 | T_v) \quad (3)$$

3.1 Web-based Estimation of Lexical Entailment Probabilities

We perform unsupervised empirical estimation of the lexical entailment probabilities, $P(Tr_u=1|T_v)$, based on word co-occurrence frequencies from the web. Following our proposed probabilistic model (cf. Section 2.1), we assume that the web is a sample generated by a language source. Each document represents a generated text and a (hidden) possible world. Given that the possible world of the text is not observed we do not know the truth assignments of hypotheses for the observed texts. We therefore further make the simplest assumption that all hypotheses stated verbatim in a document are true and all others are false and hence $P(Tr_u=1|T_v) = P(T_u|T_v)$. This simple co-occurrence probability, which we denote as lexical entailment probability – $lep(u,v)$, is easily estimated based on maximum likelihood counts:

$$lep\ (u, v) \approx P(T_u | T_v) \approx \frac{n_{u,v}}{n_v} \quad (4)$$

where n_v is the number of documents containing word v and $n_{u,v}$ is the number of documents containing both u and v. The corresponding counts were achieved by performing queries to a web search engine.

The lexical entailment probability is derived from (4) and (5) above as follows:

$$P(Tr_h = 1 | t) = \prod_{u \in h} \max_{v \in t} lep(u, v) \quad (5)$$

4 Experimental Setting

The text and hypotheses of all pairs in development set and test set where tokenized by the following simple heuristic – split at white space and remove any preceding or trailing of these characters: ([{)]}""`.,;:-!?. A stop list was applied to remove frequent tokens. Counts were obtained using the *AltaVista* search engine[2], which supplies an estimate for the number of results (web-pages) for a given one or two token query.

We empirically tuned a threshold, λ, on the the estimated entailment probability to decide if entailment holds on not. For a pair <t,h>, we tag an example as true (i.e. entailment holds) if $p= P(Tr_h = 1| t) > \lambda$, and as false otherwise. We assigned a confidence of p to the positive examples ($p>\lambda$) and a confidence of 1-p to the negative ones.

The threshold was tuned on the on the 567 annotated text-hypothesis example pairs in the development set. The optimal (best cws) threshold was λ =0.005 with a resulting cws of 0.57 and accuracy of 56%. This threshold was used to tag and assign confidence scores to the 800 pairs of the test set.

4.1 Results

The resulting accuracy on the test set was of 59% and the resulting confidence weighted score was of 0.57. Both are statistically significantly better then chance at the 0.01 level.

4.2 Analysis

Table 1 lists the accuracy and cws when computed separately for each task. As can be seen by the table the system does well on the CD and MT tasks, and quite poorly (not better than chance) on the RC, PP, IR and QA tasks.

task	accuracy	cws
Comparable Documents (CD)	0.8333	0.8727
Machine Translation (MT)	0.5667	0.6052
Information Extraction (IE)	0.5583	0.5143
Reading Comprehension (RC)	0.5286	0.5142
Paraphrase (PP)	0.5200	0.4885
Information Retrieval (IR)	0.5000	0.4492
Question Answering (QA)	0.4923	0.3736

Table 1: accuracy and cws by task

It seems as if the success of the system is attributed almost solely to its success on the CD and MT tasks. Indeed it seems as if there is something common to these two tasks, which differentiates them from the others - in both tasks high overlap of content words (or their meanings) tend to correspond to entailment.

Success and failure cases
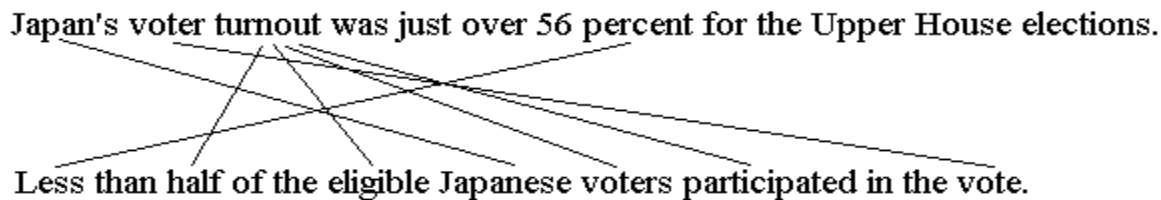The system misclassified 331 out of the 800 test examples. The vast majority of these mistakes

Figure 2: system's underlying alignment for example 1026 (RC). gold standard - false, system - false

(75%) were false positives – pairs the system tagged as true but annotated as false. It is also interesting to note that the false negative errors were more common among the MT and QA tasks while the false positive errors were more typical to the other tasks. An additional observation from the recall-precision curve (Figure 1), is that high system confidence actually corresponds to false entailment. This is attributed to an artifact of this dataset by which examples with high word overlap between the text and hypothesis tend to be biased to negative examples.
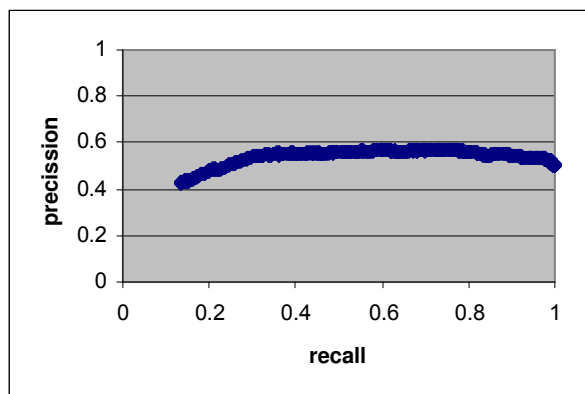

Figure 1: precision recall curve of system

In an attempt to 'look under the hood' we exmined at the underlying alignment preformed by our system on a sample of examples. Figure 2 illustrates a typical alignment. Though some of the entailing words correspond to what we believe to be the correct alignment (e.g. voter → vote, japan's → japanese), the system also finds many dubious lexical pairs (e.g. turnout → half, percent → less). Obviously, co-occurrence within documents is only one factor in estimating the entailment between words. This information should be combined with other statistical criteria that capture complementary notions of entailment, as addressed in (Geffet and

Dagan, 2004), or with lexical resources such as WordNet.

In an additional experiment we tried using as a confidence score a weighted average of the lexical probabilities (rather than the product in Equation 1) using the token's *idf* as a weight, following the weighting scheme which was applied to direct word overlap in (Monz and de Rijke, 2001). This method resulted in comparable but slightly lower accuracy of 56%.

5 Conclusions

This paper described the Bar-Ilan system participating in the Recognising Textual Entailment Challenge. We proposed a general probabilistic setting that formalizes the notion of textual entailment. In addition we described a model for lexical entailment based on web co-occurrence statistics in a bag of words representation. Although our proposed lexical system is relatively simple, as it doesn't rely on syntactic or other deeper analysis; it nevertheless achieved encouraging results. The results suggest that the proposed probabilistic framework is a promising basis for improved implementations incorporating deeper types of information.

References

Ido Dagan and Oren Glickman. 2004. *Probabilistic Textual Entailment: Generic Applied Modeling of Language Variability*. PASCAL workshop on Learning Methods for Text Understanding and Mining.

Maayan Geffet and Ido Dagan. 2004. *Feature Vector Quality and Distributional Similarity*, Coling 2004.

Christof Monz, Maarten de Rijke. 2001. Light-Weight Entailment Checking for Computational Semantics. In Proc. of the third workshop on inference in computational semantics (ICoS-3).