A Probabilistic Setting and Lexical Cooccurrence Model for Textual Entailment

Oren Glickman and Ido Dagan
Department of Computer Science
Bar Ilan University
{glikmao,Dagan}@cs.biu.ac.il

Abstract

This paper proposes a general probabilistic setting that formalizes a probabilistic notion of textual entailment. We further describe a particular preliminary model for lexical-level entailment, based on document cooccurrence probabilities, which follows the general setting. The model was evaluated on two application independent datasets, suggesting the relevance of such probabilistic approaches for entailment modeling.

1 Introduction

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), text summarization and Machine Translation (MT) evaluation are examples of applications that need to assess this semantic relationship between text segments. The Textual Entailment Recognition task (Dagan et al., 2005) has recently been proposed as an application independent framework for modeling such inferences.

Within the textual entailment framework, a text *t* is said to entail a textual hypothesis *h* if the truth of *h* can be inferred from *t*. 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 a hypothesized answer. Given the question "Does John Speak French?", a text that includes the sentence "*John is a fluent French speaker*" entails the suggested answer "John speaks French." In many cases, though, entailment inference is uncertain and has a probabilistic nature. For example, a text that includes the sentence "*John was born in France.*" does not strictly entail the above answer. Yet, it is clear that it does increase substantially the likelihood that the hypothesized answer is true.

The uncertain nature of textual entailment calls for its explicit modeling in probabilistic terms. We therefore propose a general generative probabilistic setting for textual entailment, which allows a clear formulation of concrete probabilistic models for this task. We suggest that the proposed setting may provide a unifying framework for modeling uncertain semantic inferences from texts.

An important sub task of textual entailment, which we term *lexical entailment*, is recognizing if the lexical concepts in a hypothesis *h* are entailed from a given text *t*, even if the relations which hold between these concepts may not be entailed from *t*. This is typically a necessary, but not sufficient, condition for textual entailment. For example, in order to infer from a text the hypothesis "Chrysler stock rose," it is a necessary that the concepts of *Chrysler*, *stock* and *rise* must be inferred from the text. However, for proper entailment it is further needed that the right relations hold between these concepts. In this paper we demonstrate the relevance of the general probabilistic setting for modeling lexical entailment, by devising a preliminary model that is based on document co-occurrence probabilities 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 was among the top ranking systems in the first Recognising Textual Entailment (RTE) Challenge (Glickman et al., 2005a). The model was evaluated also on an additional dataset, where it compares favorably with a state-of-the-art heuristic score. These results suggest that the proposed probabilistic framework is a promising basis for devising improved models that incorporate richer information.

Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pages 43–48,
Ann Arbor, June 2005.

example	text	hypothesis
1	*John is a French Speaker*	John speaks French
2	*John was born in France*	
3	*Harry's birthplace is Iowa*	*Harry was born in Iowa*
4	*Harry is returning to his Iowa hometown*	

Table 1: example sentence pairs

2 Probabilistic Textual Entailment

2.1 Motivation

A common definition of entailment in formal semantics (Chierchia. and McConnell-Ginet, 1990) specifies that a text t entails another text h (hypothesis, in our terminology) if h is true in every circumstance (possible world) in which t is true. For example, in examples 1 and 3 from Table 1 we'd assume humans to agree that the hypothesis is necessarily true in any circumstance for which the text is true. In such intuitive cases, textual entailment may be perceived as being certain, or, taking a probabilistic perspective, as having a probability of 1.

In many other cases, though, entailment inference is uncertain and has a probabilistic nature. In example 2, the text doesn't contain enough information to infer the hypothesis' truth. And in example 4, the meaning of the word *hometown* is ambiguous and therefore one cannot infer for certain that the hypothesis is true. In both of these cases there are conceivable circumstances for which the text is true and the hypothesis false. Yet, it is clear that in both examples, the text does increase substantially the likelihood of the correctness of the hypothesis, which naturally extends the classical notion of certain entailment. Given the text, we expect the probability that the hypothesis is indeed true to be relatively high, and significantly higher than its probability of being true without reading the text. Aiming to model application needs, we suggest that the probability of the hypothesis being true given the text reflects an appropriate confidence score for the correctness of a particular textual inference. In the next subsections we propose a concrete probabilistic setting that formalizes the notion of truth probabilities in such cases.

2.2 A Probabilistic Setting

Let T denote a space of possible texts, and $t \in T$ a specific text. 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 it could also be expressed as a formula in some propositional language.

A semantic state of affairs is captured by a mapping from H to {0=false, 1=true}, denoted by $w: H \rightarrow \{0, 1\}$ (called here *possible world*, following common terminology). A possible world w represents a concrete set of truth value assignments for all possible propositions. Accordingly, W denotes the set of all possible worlds.

2.2.1 A Generative Model

We assume a probabilistic generative model for texts and possible worlds. In particular, we assume that texts are generated along with a concrete state of affairs, represented by a possible world. Thus, whenever the source generates a text t, it generates also corresponding hidden truth assignments that constitute a possible world w.

The probability distribution of the source, over all possible texts and truth assignments $T \times W$, is assumed to reflect 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 a particular "real" world, but only the inferences about propositions' truth which 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 $P(h)$. 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 or France. In fact, we may as well assume that the notion of textual entailment is relevant only for hypotheses for which $P(h) < 1$, as otherwise (i.e. for tautologies) there is no need to consider texts that would support h's truth. 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 a given world. Correspondingly, Tr_h=1 is the event of h being assigned a truth value of 1 (true).
II) For a text t, we use t itself 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).

2.3 Probabilistic textual entailment definition

We say that a text t probabilistically entails a hypothesis h (denoted as $t \Rightarrow h$) if t increases the likelihood of h being true, that is, if $P(Tr_h = 1| t) > P(Tr_h = 1)$ or equivalently if the pointwise mutual information, I(Trh=1,t), is greater then 0. Once *knowing* that $t \Rightarrow h$, $P(Tr_h=1| t)$ serves as a probabilistic confidence value for h being true given t.

Application settings would typically require that $P(Tr_h = 1| t)$ obtains a high value; otherwise, the text would not be considered sufficiently relevant to support h's truth (e.g. a supporting text in QA or IE should entail the extracted information with high confidence). Finally, we ignore here the case in which t contributes negative information about h, leaving this relevant case for further investigation.

2.4 Model Properties

It is interesting to notice the following properties and implications of our model:
A) Textual entailment is defined as a relationship between texts and propositions whose representation is typically based on text as well, unlike logical entailment which is a relationship between propositions only. Accordingly, textual entailment confidence is conditioned on the actual generation of a text, rather than its truth. For illustration, we would expect that the text "His father was born in Italy" would logically entail the hypothesis "He was born in Italy" with high probability – since most people who's father was born in Italy were also born there. However we expect that the text would actually not probabilistically textually entail the hypothesis since most people for whom it is specifically reported that their father was born in Italy were not born in Italy.[1]
B) We assign probabilities to propositions (hypotheses) in a similar manner to certain probabilistic reasoning approaches (e.g. Bacchus, 1990; Halpern, 1990). However, we also assume a generative model of text, similar to probabilistic language and machine translation models, which supplies the needed conditional probability distribution. Furthermore, since our conditioning is on texts rather than propositions we do not assume any specific logic representation language for text meaning, and only assume that textual hypotheses can be assigned truth values.
C) Our framework does not distinguish between textual entailment inferences that are based on knowledge of language semantics (such as *murdering* $\Rightarrow$ *killing*) and inferences based on domain or world knowledge (such as *live in Paris* $\Rightarrow$ *live in France*). Both are needed in applications and it is not clear at this stage where and how to put such a borderline.
D) An important feature of the proposed framework is that for a given text many hypotheses are likely to be true. Consequently, for a given text t and hypothesis h, $\sum_h P(Tr_h=1|t)$ does not sum to 1. This differs from typical generative settings for IR and MT (Ponte and croft, 1998; Brown et al., 1993), where all conditioned events are disjoint by construction. In the proposed model, it is rather the case that $P(Tr_h=1|t) + P(Tr_h=0|t) = 1$, as we are interested in the probability that a single particular hypothesis is true (or false).
E) An implemented model that corresponds to our probabilistic setting is expected to produce an estimate for $P(Tr_h = 1| t)$. This estimate is expected to reflect all probabilistic aspects involved in the modeling, including inherent uncertainty of the entailment inference itself (as in example 2 of Table 1), possible uncertainty regarding the correct disambiguation of the text (example 4), as well as probabilistic estimates that stem from the particular model structure.

3 A Lexical Entailment Model

We suggest that the proposed setting above provides the necessary grounding for probabilistic

[1] This seems to be the case, when analyzing the results of entering the above text in a web search engine.

modeling of textual entailment. Since modeling the full extent of the textual entailment problem is clearly a long term research goal, in this paper we rather focus on the above mentioned sub-task of *lexical entailment* - identifying when the lexical elements of a textual hypothesis *h* are inferred from a given text *t*.

To model lexical entailment we first assume that the meanings of the individual content words in a hypothesis can be assigned truth values. One possible interpretation for such truth values is that lexical concepts are assigned existential meanings. For example, for a given text *t*, $Tr_{book}=1$ if it can be inferred in *t*'s state of affairs that a book exists. Our model does not depend on any such particular interpretation, though, as we only assume that truth values can be assigned 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 as well. This captures our target perspective of lexical entailment, while not modeling here other entailment aspects. When estimating the entailment probability we assume that the truth probability of a term *u* in a hypothesis *h* is independent of the truth of the other terms in *h*, obtaining:

$$P(Tr_h = 1 | t) = \Pi_{u \in h} P(Tr_u = 1 | t)$$
$$P(Tr_h = 1) = \Pi_{u \in h} P(Tr_u = 1) \quad (1)$$

In order to estimate $P(Tr_u=1|v_1, \ldots, v_n)$ 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 *Tv* 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[2]). Alternatively, one can view (2) as inducing an alignment between terms in the *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:

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

[2] Such a model is proposed in (Glickman et al., 2005b)

3.1 Estimating 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 in a corpus. Following our proposed probabilistic model (cf. Section 2.2.1), we assume that the domain corpus 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 from the corpus based on maximum likelihood counts:

$$lep(u, v) = P(Tr_u = 1 | T_v) \approx \frac{n_{u,v}}{n_v} \quad (4)$$

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*.

Given our definition of the textual entailment relationship (cf. Section 2.3) for a given word *v* we only consider for entailment words *u* for which $P(Tr_u=1|T_v) > P(Tr_u=1)$ or based on our estimations, for which $n_{u,v}/n_u > n_v/N$ (N is total number of documents in the corpus).

We denote as *tep* the textual entailment probability estimation as derived from (3) and (4) above:

$$tep(t, h) = \prod_{u \in h} \max_{v \in t} lep(u, v) \quad (5)$$

3.2 Baseline model

As a baseline model for comparison, we use a score developed within the context of text summarization. (Monz and de Rijke, 2001) propose modeling the directional entailment between two texts t_1, t_2 via the following score:

$$entscore(t_1, t_2) = \frac{\sum_{w \in (t_1 \cap t_2)} idf(w)}{\sum_{w \in t_2} idf(w)} \quad (6)$$

where $idf(w) = \log(N/n_w)$, N is total number of documents in corpus and n_w is number of docu-

ments containing word *w*. A practically equivalent measure was independently proposed in the context of QA by (Saggion et al., 2004)[3]. This baseline measure captures word overlap, considering only words that appear in both texts and weighs them based on their inverse document frequency.

4 The RTE challenge dataset

The RTE dataset (Dagan et al., 2005) consists of sentence pairs annotated for entailment. Fo this dataset we used word cooccurrence frequencies obtained from a web search engine. The details of this experiment are described in Glickman et al., 2005a. The resulting accuracy on the test set was 59% and the resulting confidence weighted score was 0.57. Both are statistically significantly better than chance at the 0.01 level. The baseline model (6) from Section 3.2, which takes into account only terms appearing in both the text and hypothesis, achieved an accuracy of only 56%. Although our proposed lexical system is relatively simple, as it doesn't rely on syntactic or other deeper analysis, it nevertheless was among the top ranking systems in the RTE Challenge.

5 RCV1 dataset

In addition to the RTE dataset we were interested in evaluating the model on a more representative set of texts and hypotheses that better corresponds to applicative settings. We focused on 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.

An annotator was asked to choose 60 hypotheses based on sentences from the first few documents in the *Reuters Corpus Volume 1* (Rose et al., 2002). The annotator was instructed to choose sentential hypotheses such that their truth could easily be evaluated. We further 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. Table 2 shows a few of the hypotheses.

In order to create a set of candidate entailing texts for the given set of test hypotheses, we followed the common practice of WordNet based expansion (Nie and Brisebois, 1996; Yang and Chua, 2002). Using WordNet, we expanded the hypotheses' terms with morphological alternations and semantically related words[4].

For each hypothesis stop words were removed and all content words were expanded as described above. Boolean Search included a conjunction of the disjunction of the term's expansions and was performed at the paragraph level over the full Reuters corpus, as common in IR for QA. Since we wanted to focus our research on semantic variability we excluded from the result set paragraphs that contain all original words of the hypothesis or their morphological derivations. The resulting dataset consists of 50 hypotheses and over a million retrieved paragraphs (10 hypotheses had only exact matches). The number of paragraphs retrieved per hypothesis range from 1 to 400,000.[5]

5.1 Evaluation

The model's entailment probability, *tep,* was compared to the following two baseline models. The first, denoted as *base*, is the naïve baseline in which all retrieved texts are presumed to entail the hypothesis with equal confidence. This baseline corresponds to systems which perform blind expansion with no weighting. The second baseline, *entscore*, is the entailment score (6) from 3.2.

The top 20 best results for all methods were given to judges to be annotated for entailment. Judges were asked to annotate an example as true if given the text they can infer with high confidence that the hypothesis is true (similar to the guidelines published for the RTE Challenge dataset). Accordingly, they were instructed to annotate the example as false if either they believe the hypothesis is false given the text or if the text is unrelated to the hypothesis. In total there were 1683 text-hypothesis pairs, which were randomly divided between two judges. In order to measure agreement, we had 200 of the pairs annotated by both judges, yielding a moderate agreement (a *Kappa* of 0.6).

[3] (Saggion et al., 2004) actually proposed the above score with no normalizing denominator. However for a given hypothesis it results with the same ranking of candidate entailing texts.

[4] The following WordNet relations were used: *Synonyms*, *see also*, *similar to, hypernyms/hyponyms*, *meronyms/holonyms*, *pertainyms*, *attribute*, *entailment*, *cause* and *domain*

[5] The dataset is available at: http://ir-srv.cs.biu.ac.il:64080/emsee05_dataset.zip

5.2 Results

	base	entscore	tep
precision	0.464	0.568	0.647
cws	0.396	0.509	0.575

Table 2: Results

Table 2 includes the results of macro averaging the precision at top-20 and the average *confidence weighted score* (cws) achieved for the 50 hypotheses. Applying Wilcoxon Signed-Rank Test, our model performs significantly better (at the 0.01 level) than entscore and base for both precision and cws. Analyzing the results showed that many of the mistakes were not due to wrong expansion but rather to a lack of a deeper analysis of the text and hypothesis (e.g. example 3 in Table 2). Indeed this is a common problem with lexical models. Incorporating additional linguistic levels into the probabilistic entailment model, such as syntactic matching, co-reference resolution and word sense disambiguation, becomes a challenging target for future research.

6 Conclusions

This paper proposes a generative probabilistic setting that formalizes the notion of probabilistic textual entailment, which is based on the conditional probability that a hypothesis is true given the text. This probabilistic setting provided the necessary grounding for a concrete probabilistic model of lexical entailment that is based on document co-occurrence statistics in a bag of words representation. Although the illustrated 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 such a probabilistic framework is a promising basis for improved implementations incorporating deeper types of knowledge and a common test-bed for more sophisticated models.

Acknowledgments

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. We would also like to thank Ruthie Mandel and Tal Itzhak Ron for their annotation work.

References

Fahiem Bacchus. 1990. *Representing and Reasoning with Probabilistic Knowledge*, M.I.T. Press.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics, 19(2):263–311.

Chierchia, Gennaro, and Sally McConnell-Ginet. 2001. *Meaning and grammar: An introduction to semantics*, 2nd. edition. Cambridge, MA: MIT Press.

Ido Dagan, Oren Glickman and Bernardo Magnini. 2005. *The PASCAL Recognising Textual Entailment Challenge*. In Proceedings of the PASCAL Challenges Workshop for Recognizing Textual Entailment. Southampton, U.K.

Oren Glickman, Ido Dagan and Moshe Koppel. 2005a. *Web Based Probabilistic Textual Entailment*, PASCAL Challenges Workshop for Recognizing Textual Entailment.

Oren Glickman, Ido Dagan and Moshe Koppel. 2005b. *A Probabilistic Classification Approach for Lexical Textual Entailment*, Twentieth National Conference on Artificial Intelligence (AAAI-05).

Joseph Y. Halpern. 1990. *An analysis of first-order logics of probability. Artificial Intelligence* 46:311-350.

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).

Jian-Yun Nie and Martin Brisebois. 1996. *An Inferential Approach to Information Retrieval and Its Implementation Using a Manual Thesaurus*. Artificial Intelligence Revue 10(5-6): 409-439.

Jay M. Ponte, W. Bruce Croft, 1998. *A Language Modeling Approach to Information Retrieval*. SIGIR conference on Research and Development in Information Retrieval.

Tony G. Rose, Mary Stevenson, and Miles Whitehead. 2002. *The Reuters Corpus volume 1 - from yesterday's news to tomorrow's language resources*. Third International Conference on Language Resources and Evaluation (LREC).

Hui Yang and Tat-Seng Chua. 2002. *The integration of lexical knowledge and external resources for question answering*. The eleventh Text REtrieval Conference (TREC-11).