A Probabilistic Classification Approach for Lexical Textual Entailment

Oren Glickman and **Ido Dagan** and **Moshe Koppel**

Computer Science Department , Bar Ilan University
Ramat-Gan, 52900, Iarael
{glikmao, dagan, koppel}@cs.biu.ac.il

Abstract

The textual entailment task – determining if a given text entails a given hypothesis – provides an abstraction of applied semantic inference. 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.

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 such semantic relationship between text segments. *Textual Entailment Recognition* (Dagan et al., 2005) has recently been proposed as an application independent task to capture 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. In this paper we first propose a general generative probabilistic setting for textual entailment. This probabilistic framework may be considered analogous to (though different than) the probabilistic setting defined for other phenomena, like language modeling and statistical machine translation. Obviously, such frameworks are needed to allow for principled (fuzzy) modeling of the relevant language phenomenon, compared to utilizing ad-hoc ranking scores which have no clear probabilistic interpretation. 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 concepts in a hypothesis *h* are entailed from a given text *t*, even if the relations 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 "Yahoo acquired Overture," it is necessary that the concepts of *Yahoo*, *acquisition* and *Overture* must all be inferred from the text. However, for proper entailment it is further needed that the right relations would hold between these concepts.

Next, we demonstrate the utility of our probabilistic setting by developing a rather simple concrete probabilistic model for the lexical entailment sub-problem, which follows the general setting. The model recasts the lexical entailment problem as a variant of text categorization, where classes correspond to all content words, which represent hypothesis concepts. Recognizing whether a hypothesis concept (content word) is entailed from the text *t* is carried out by classifying *t* to the corresponding class. To this end we utilize a Naïve Bayes classifier in a novel and completely unsupervised fashion, based on crude raw estimates.

Empirical tests were conducted by simulating common lexical expansion based on WordNet. Results indicate that the simple model is effective and compares favorably with a baseline of a typical state-of-the-art scoring function.

Background

Dealing with semantic inferences between texts is a common problem in NLP. Within application settings a wide variety of semantic inference techniques were proposed, ranging at different levels of representation and complexity. Various works (Moldovan and Rus 2001; Hobbs et al. 1993; Condoravdi et al. 2003) utilize deep semantic representations by interpreting text into a formal language on which inference is performed. However, such techniques are not commonly used due to their computational costs and the lack of sufficiently efficient and robust NLP tools. Furthermore, such approaches are often geared for deterministic inferences and less so for the uncertain type of reasoning addressed here.

Recent works on paraphrase and entailment rule acquisition (Barzilay and McKeown 2001; Lin and Pantel 2001; Szpektor et al. 2004) learn lexical syntactic paraphrase patterns. Though clearly related to our problem, this work usually ignores the directionality aspect of entailment, and produces heuristic scores which are not clearly usable for inference. It would be appealing to develop probabilistic models analogous to our lexical one which apply to these more complex patterns.

Within the lexical scope, WordNet-based term expansion is the most commonly used technique for enhancing the recall of NLP systems and coping with lexical variability (e.g. Califf and Mooney, 2003; Nie and Brisebois. 1996). For example, many QA systems perform a lexical expansion retrieval phase, which returns a relatively small ordered set of candidate answer-bearing texts. In some QA systems, (Harabagiu et al., 2000; Kwok et al., 2001; Hovy et al., 2001) there is no clear weighting scheme for the lexical expansion, and expanded words are added equally to a Boolean retrieval search of candidate answer texts. Saggion et al. (2004) do propose ranking the candidate answer texts for a given question based on an idf-weighted measure for the degree of word overlap between the question and the candidate test, as follows: $\mathrm{score}(t,h)=\Sigma_{u\in t\cap h}\,\mathrm{idf}(u)$. This measure favors texts that were retrieved by fewer expansions or by expansions of relatively frequent words. An equivalent measure was proposed in the context of summarization (Monz and de Rijke 2001) to specifically model directional entailment between texts.

In summary, many approaches for modeling entailment were developed in application specific settings. Furthermore, in an abstract application-independent setting it is not clear how scores for semantic variations should be assigned and interpreted, which may call for a generic probabilistic setting for textual entailment.

Probabilistic Textual Entailment

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, given the hypothesis h_1 = "*Marry Speaks French*" and a candidate text t_1 that includes the sentence "*Marry is a Fluent French Speaker*", it is clear that t_1 strictly entails h_1, and humans are likely to have high agreement regarding this decision. In many other cases, though, entailment inference is uncertain and has a probabilistic nature. For example, a text t_2 that includes the sentence "*Marry was born in France.*" does not strictly entail the above h_1 (i.e. there are circumstances in which someone was born in France but yet doesn't speak French). 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. Thus, in this example, 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 significantly higher than its probability of being true without reading the text. In the next section we propose a concrete probabilistic setting that formalizes the notion of truth probabilities in such cases.

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 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 $w\colon H \rightarrow \{0, 1\}$ (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.

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.

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). 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 the world of the generated text. 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 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, i.e. $P(\mathrm{Tr}_h=1| t)>P(\mathrm{Tr}_h=1)$, or equivalently if the pointwise mutual information, $I(\mathrm{Tr}_h=1,t)$, is greater then 1. Once knowing that $t \Rightarrow h$, $P(\mathrm{Tr}_h=1| t)$ serves as a probabilistic confidence value for h being true given t.

An Unsupervised Lexical Model

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 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. It is important to bear in mind that it is not trivial to estimate the constituent probabilities which correspond to the textual entailment framework since the possible worlds of texts are not observed and we do not know the corresponding truth assignments of hypotheses.

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, $\mathrm{Tr}_{\mathrm{acquired}}$=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 as well (capturing 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 in a hypothesis h is independent of the truth of the other terms in h, obtaining:

$$\begin{aligned} P(\mathrm{Tr}_h = 1| t) &= \Pi_{u \in h} P(\mathrm{Tr}_u=1|t) \\ P(\mathrm{Tr}_h = 1) &= \Pi_{u \in h} P(\mathrm{Tr}_u=1) \end{aligned} \quad (1)$$

Textual Entailment as Text Classification

At this point, it is perhaps best to think of the entailment problem as a text classification task. Our main sleight-of-hand here is estimating these probabilities, $P(\mathrm{Tr}_u = 1| t)$ for text t and a lexical item u as text classification probabilities in which the classes are the different words u in the vocabulary. Following this perspective we apply a technique commonly used for text classification. Our proposed model resembles work done on text classification from labeled and unlabeled examples (Nigam et al., 2000), using a Naïve Bayes classifier. However, our setting and task is quite different – we classify texts to a binary abstract notion of lexical truth rather than to well-defined supervised classes. We utilize unsupervised initial approximate labeling of classes, while Nigam et al. bootstrap from labeled data. First, we construct the initial labeling based solely on the explicit presence or absence of each u in t. Then we apply Naïve Bayes in an unsupervised fashion derived 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. Thus, referring to the given corpus texts at the document level, we have $P(\mathrm{Tr}_u = 1| t)=1$ if $u \in t$ and 0 otherwise, which defines the initial class labels for every term u and text t (training labels). It also follows from (1) that a text entails a hypothesis if and only if it contains all content words of the hypothesis.

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 (some examples of this can be seen in Table 1), which is addressed in the subsequent model. The initial labeling, however, may provide useful estimates for this model.

Naïve Bayes Refinement

Based on the initial labeling we consider during training all texts that include u as positive examples for this class and take all the other texts as negative examples.

For a word *u,* P(Tr$_u$=1| *t*) can be rewritten, by following the standard naïve Bayes assumption, as in (2):

$$P(\mathrm{Tr}_u = 1 \mid t) = \frac{P(t \mid \mathrm{Tr}_u = 1)P(\mathrm{Tr}_u = 1)}{P(t)} =$$

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

$$* \textit{naive bayes assumption} \qquad (2)$$

$$\frac{P(\mathrm{Tr}_u = 1)\prod_{i=1}^{|t|} P(t_i \mid \mathrm{Tr}_u = 1)}{\sum_{c \in \{0,1\}} P(\mathrm{Tr}_u = c)\prod_{i=1}^{|t|} P(t_i \mid \mathrm{Tr}_u = c)}$$

where t_i is the *i*'th word of *t*.

In this way we are able to estimate P(Tr$_u$=1| *t*) based solely on the prior probabilities P(Tr$_u$=1), P(Tr$_u$=0) and the lexical co-occurrence probabilities P(*v*| Tr$_u$=1), P(*v*| Tr$_u$=0) for *u*, *v* in the vocabulary *V*. These probabilities are easily estimated from the corpus given the initial model's estimate of truth assignments. Estimation is done assuming a multinomial event model for documents and Laplace smoothing (McCallum and Nigam 1998) as in (3):

$$P(\mathrm{Tr}_u = 1) = \frac{|d \in D : u \in d|}{|D|}$$

$$P(v \mid \mathrm{Tr}_u = 1) = \frac{1 + \sum_{d \in D : u \in d} N(v,d)}{|V| + \sum_{v' \in V}\sum_{d \in D : u \in d} N(v',d)} \qquad (3)$$

where *v* and *u* are words, *D* is the set of documents in the corpus, and *N*(*v*,*d*) is the occurrence frequency of *v* in *d*.

From equations (1), (2) and (3) we have a refined probability estimate for P(Tr$_h$ = 1| *t*) and P(Tr$_h$ = 1) for any arbitrary text *t* and hypothesis *h* (estimation for the case of Tr$_u$=0 is analogous).

The criterion for turning probability estimates into classification decisions is derived analytically from our proposed probabilistic setting of textual entailment. We make a positive classification for entailment if P(Tr$_h$ = 1| *t*) > P(Tr$_h$ = 1) and assign a confidence score of P(Tr$_h$ = 1| *t*) for ranking purposes. In fact, the empirical evaluation showed this analytic threshold to be almost optimal.

Empirical Evaluation

Experimental Setup

Though empirical modeling of semantic inferences between texts is commonly done within application settings, we wanted to specifically evaluate a textual entailment system in an application independent manner. 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* (Rose et al. 2002) - a collection of about 810,000 English News stories most of which are economy related. An annotator chose 50 hypotheses based on sentences from the first few documents in the Reuters corpus. The annotator was instructed to choose short 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. These annotations were used solely for the evaluation process since our proposed method is unsupervised and does not rely on annotations. A few of the hypotheses can be seen in Table 1.

In order to create a plausible set of candidate entailing texts for the given set of test hypotheses, we followed the common practice of morphological and WordNet-based expansion. For each hypothesis, stop words were first removed and all content words were expanded using WordNet's morphological alternations and semantically related words[1]. Expansion was done for each possible sense for each word type in the hypothesis. Boolean search was then performed at the paragraph level over the full Reuters corpus. The Boolean query includes a conjunction of the disjunction of the terms' expansions.

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. We then picked a random set of 20 texts for each of the 50 hypotheses. The resulting dataset was given to two judges to be annotated for entailment. Corresponding to our generic notion of textual entailment as defined in section 2, judges were asked to annotate a text-hypothesis pair as true if, given the text, they could infer with high 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[2]. Overall, the annotators deemed 48% of the text-hypothesis pairs as positive examples of entailment. Note that although our model is targeted on identifying lexical entailment we are judging its effectiveness on the general textual entailment task.

In order to assess agreement, a small subset of 200 pairs was cross annotated. This resulted in a moderate Kappa statistic of 0.6 (Landis and Koch, 1977). The relatively low (but still significant) Kappa value may be attributed to the probabilistic nature of the defined task, but might still be improved in the future through improved judgment guidelines and practices.

[1] The following relations were used, based on common practice in the literature: *Synonyms*, *cause, pertainyms, meronyms/holonyms*, *hypernyms/hyponyms, similar to*, *attribute*, *see also*, and *domain*

[2] The annotated dataset is available at: http://ir-srv.cs.biu.ac.il:64080/aaai05_dataset.zip

#	text	hypothesis	system	judge
1	Wall Street ended a stormy session with sharp losses Wednesday after the stock market surrendered a rally after news that the Federal Reserve was keeping interest rates unchanged.	federal reserve raise interest rates	1	0
2	The Libyan officials, accompanied by members of a Palestinian youth group, arrived at the camp in several small buses but it was not clear how they intended to transport the residents of the camp.	assemble truck	0	0
3	Earlier Tuesday CompuServe reported on its first quarter, warned of the expected second-quarter loss and said it expects improvements in the second half.	CompuServe predicted a loss	1	1
4	Fresh signs of labor market tightness emerged from the August employment report released Friday morning. The economy created 250,000 new jobs, while the unemployment rate fell to 5.1 percent, a six-and-a-half-year low.	cut jobs	1	0
5	Tokyo investors couldn't wait to get back to business on Tuesday after a long weekend, sending the key Nikkei stock average up by more than two percent, on the back of Wall Street's record climb in overnight trade.	Nikkei average rose	1	1
6	Shares in tobacco group Seita, privatised last year, slipped on Thursday in a quiet market under the impact of negative publicity concerning a second civil suit in a week by the family of smokers who died.	tobacco industry sued	0	1
7	Payne also asked whether Lloyd's was able to pay the total premium if U.S. investors did not back the recovery plan. Sandler again said he was unable to answer.	Sandler questioned	0	1
8	"We're starting to see some growth. We're seeing a recovery in fees and we can view that as a leading indicator of growth in the economy," said Susana Ornelas, banking analyst at Deutsche Morgan Grenfell in Mexico City.	Mexico's economy is recovering	1	1

Table 1: text-hypothesis examples along with classification results and annotation

Classification Results

We trained our model on the Reuters corpus and then classified the test text-hypothesis pairs. The model's predictions were then compared with the human judgments. The resulting average accuracy per hypothesis (macro averaging) was of 70%. Since scoring semantic variations is commonly treated as a ranking task (see Background), there is no clear baseline system for accuracy comparison.

The straw baseline of predicting all pairs to be false (i.e. the text does not entail the hypothesis) corresponds to our initial labeling since our dataset did not include texts containing all content words of the hypotheses. This baseline yields an average accuracy of only 52%. In order to get a rough estimation of the upper bound for accuracy for lexical based expansion models we performed an additional experiment of judging a sample of the texts in which all hypothesis content words do appear in the text (recall that these examples where excluded from our evaluation set). In this set of exact Boolean matches, on average only 82% of the texts (per hypothesis) were judged as entailing. In an additional experiment, it turns out that the natural classification threshold as derived from our probabilistic framework is almost optimal – the best threshold attains an accuracy of 71%.

Ranking Results

We also compared our system's ranking ability, particularly since most state-of-the-art approaches may provide a score but not a clear classification criterion. The entailment confidence score was used to rank the various texts of each hypothesis. The average *confidence weighted score* (*cws,* also termed average precision) was measured for each hypothesis and is calculated as follows:

$$cws = \frac{1}{N}\sum_{i=1}^{N}\frac{\#\ \text{correct up to rank } i}{i} \qquad (4)$$

where N is the number of texts for the given hypothesis.

Table 2 shows the resulting cws macro averages. The *min* and *max* cws are the lower and upper bound values for cws on this dataset (a max cws corresponds to perfect ordering). The *rand* column corresponds to a random ordering. *idf* corresponds to the state of the art procedure for weighting expansions based on word overlap of (Saggion et al. 2004; Monz and de Rijke 2001), as described in the background. This method provides a representative baseline which measures the degree of lexical overlap between the text and hypothesis, weighed by inverse document frequency. Though our model performs just somewhat better, the results are statistically significant at the 0.02 level.

	min	rand	idf	model	max
cws	0.35	0.49	0.51	0.54	0.63

Table 2: results of confidence weighted score

Analysis

Table 1 shows example text-hypothesis pairs along with their classification and judgment. 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. examples 1 and 4). Indeed this is a common problem with lexical models, which is beyond the scope of this paper. Tackling this issue is a challenging target for future research.

Our model does seem, as expected, to make good topical distinctions. For example, in example 2 from Table 1, the text does contain the words *group* and *transport* which are possible expansions of *assemble* and *truck* respectively (via a hypernymy relation - the corresponding WordNet senses are included in Table 3).

sense	gloss
assemble#v#2	collect in one place
group#v#2	form a group or group together
transport#v#1	move something or somebody around; usually over long distances
truck#v#1	convey (goods etc.) by truck

Table 3: WordNet glosses for words in expansion

However, these possible expansions are erroneous given the context of this particular example due to incorrect senses and part of speech. The proposed model correctly identifies this as a negative example without performing explicit part-of-speech tagging or word sense disambiguation. Furthermore, in examples 3, 5 and 8 the model makes correct predictions that seem to be quite hard to achieve from more sophisticated methods involving deeper semantic and syntactic representations and inference mechanisms.

A natural extension of our model would have been to apply additional refinement steps via the EM algorithm (as done in (Nigam et al. 2000). However, we checked the effect of overlaying EM on our approach and found that further iterations made no significant improvement. Another natural extension to our model is to avoid the lexical independence assumption of (1). In this approach, the initial model assumes that a hypothesis is true in a document's world if and only if it contains a paragraph in which all the hypothesis' content words appear. Note that there need not always be such documents for an arbitrary hypothesis. Nevertheless, such a model actually did not perform better than our proposed model, possibly due to data sparseness when looking for co-occurrences of all the hypothesis words. Finally, an additional interesting finding is that comparable results were achieved by our model when using only 1/8 of the corpus (100,000 documents).

Conclusions and Future Work

This paper first presented a generative probabilistic setting for textual entailment. Then, a concrete model at the lexical level was proposed, which demonstrates that the problem can be practically approached within the proposed framework. Our model casts the entailment problem as a variant of text classification, while estimating class probabilities in an unsupervised manner. Evaluations demonstrate favorable performance relative to state-of-the-art common practice, suggesting a principled probabilistic interpretation. We propose that this framework can be utilized further in future work to gradually address the additional aspects of the textual entailment phenomenon.

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 like to thank Ruthie Mandel and Tal Itzhak Ron for their help with the annotation work.

References

Barzilay, R. and McKeown, K. 2001. Extracting Paraphrases from a Parallel Corpus. *ACL*, pp. 50-57.

Califf, M. E. and Mooney, R. J. 2003. Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction. *Journal of Machine Learning Research*, 4, pp. 177-210.

Chierchia, G, and McConnell-Ginet, S. 2001. *Meaning and grammar: An introduction to semantics*, MIT Press.

Condoravdi, C.; Crouch, D.; de Paiva, V.; Stolle, R.; Bobrow, D. G. 2003. Entailment, Intensionality and Text Understanding. *HLT-NAACL 2003 Workshop on Text Meaning*.

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.

Harabagiu, S. M.; Moldovan, D. I.; Pasca, M.; Mihalcea, R.; Surdeanu, M.; Bunescu, R. C.; Girju, R.; Rus, V.; Morarescu, P. 2000. FALCON: Boosting Knowledge for Answer Engines. *The ninth Text REtrieval Conference (TREC-9)*.

Hobbs, J. R.; Stickel, M. E.; Appelt, D. E.; Martin, P. 1993. Interpretation as Abduction. *Artificial Intelligence* 63: 69-142.

Hovy, E. H.; Hermjakob, U.; Lin. 2001. The Use of External Knowledge in Factoid QA. *TREC-10*.

Landis, R. J. and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33:159-174.

Lin, D. and Pantel, P. 2001. Discovery of Inference Rules for Question Answering. *Natural Language Engineering* 7:4.

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

Moldovan, D. I. and Rus, V. 2001. Logic Form Transformation of WordNet and its Applicability to Question Answering. *ACL*.

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

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

Nigam, K; McCallum, A.; Thrun, S.; Mitchell, T. 2000. *Text Classification from Labeled and Unlabeled Documents using EM. Machine Learning*, 39(2/3). pp. 103-134.

Rose, T. G.; Stevenson, M.; Whitehead, M. 2002. The Reuters Corpus volume 1 - from yester-day's news to tomorrow's language resources. *Third International Conference on Language Resources and Evaluation (LREC-02)*.

Horacio Saggion, Rob Gaizauskas, Mark Hepple, Ian Roberts and Mark A. Greenwood. 2004. Exploring the Performance of Boolean Retrieval Strategies For Open Domain Question Answering. *In Proceedings of the Information Retrieval for Question Answering (IR4QA) Workshop at SIGIR 2004*.

Szpektor, I; Tanev, H; Dagan, I; Coppola B. 2004. Scaling Web-based Acquisition of Entailment Relations. *EMNLP*.