A Probabilistic Modeling Framework for Lexical Entailment

Eyal Shnarch
Computer Science Department
Bar-Ilan University
Ramat-Gan, Israel
shey@cs.biu.ac.il

Jacob Goldberger
School of Engineering
Bar-Ilan University
Ramat-Gan, Israel
goldbej@eng.biu.ac.il

Ido Dagan
Computer Science Department
Bar-Ilan University
Ramat-Gan, Israel
dagan@cs.biu.ac.il

Abstract

Recognizing entailment at the lexical level is an important and commonly-addressed component in textual inference. Yet, this task has been mostly approached by simplified heuristic methods. This paper proposes an initial probabilistic modeling framework for lexical entailment, with suitable EM-based parameter estimation. Our model considers prominent entailment factors, including differences in lexical-resources reliability and the impacts of transitivity and multiple evidence. Evaluations show that the proposed model outperforms most prior systems while pointing at required future improvements.

1 Introduction and Background

Textual Entailment was proposed as a generic paradigm for applied semantic inference (Dagan et al., 2006). This task requires deciding whether a textual statement (termed the *hypothesis*-H) can be inferred (entailed) from another text (termed the *text*-T). Since it was first introduced, the six rounds of the Recognizing Textual Entailment (RTE) challenges[1], currently organized under NIST, have become a standard benchmark for entailment systems.

These systems tackle their complex task at various levels of inference, including logical representation (Tatu and Moldovan, 2007; MacCartney and Manning, 2007), semantic analysis (Burchardt et al., 2007) and syntactic parsing (Bar-Haim et al., 2008; Wang et al., 2009). Inference at these levels usually requires substantial processing and resources (e.g. parsing) aiming at high performance.

Nevertheless, simple entailment methods, performing at the *lexical* level, provide strong baselines which most systems did not outperform (Mirkin et al., 2009; Majumdar and Bhattacharyya, 2010). Within complex systems, lexical entailment modeling is an important component. Finally, there are cases in which a full system cannot be used (e.g. lacking a parser for a targeted language) and one must resort to the simpler lexical approach.

While lexical entailment methods are widely used, most of them apply ad hoc heuristics which do not rely on a principled underlying framework. Typically, such methods quantify the degree of lexical *coverage* of the hypothesis terms by the text's terms. Coverage is determined either by a direct match of identical terms in T and H or by utilizing lexical semantic resources, such as WordNet (Fellbaum, 1998), that capture lexical entailment relations (denoted here as entailment *rule*s). Common heuristics for quantifying the degree of coverage are setting a threshold on the percentage coverage of H's terms (Majumdar and Bhattacharyya, 2010), counting absolute number of uncovered terms (Clark and Harrison, 2010), or applying an Information Retrieval-style vector space similarity score (MacKinlay and Baldwin, 2009). Other works (Corley and Mihalcea, 2005; Zanzotto and Moschitti, 2006) have applied a heuristic formula to estimate the similarity between text fragments based on a similarity function between their terms.

These heuristics do not capture several important aspects of entailment, such as varying reliability of

[1] http://www.nist.gov/tac/2010/RTE/index.html

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 558–563,
Portland, Oregon, June 19-24, 2011.

entailment resources and the impact of rule chaining and multiple evidence on entailment likelihood. An additional observation from these and other systems is that their performance improves only moderately when utilizing lexical resources[2].

We believe that the textual entailment field would benefit from more principled models for various entailment phenomena. Inspired by the earlier steps in the evolution of Statistical Machine Translation methods (such as the initial IBM models (Brown et al., 1993)), we formulate a concrete generative probabilistic modeling framework that captures the basic aspects of lexical entailment. Parameter estimation is addressed by an EM-based approach, which enables estimating the hidden lexical-level entailment parameters from entailment annotations which are available only at the sentence-level.

While heuristic methods are limited in their ability to wisely integrate indications for entailment, probabilistic methods have the advantage of being extendable and enabling the utilization of well-founded probabilistic methods such as the EM algorithm.

We compared the performance of several model variations to previously published results on RTE data sets, as well as to our own implementation of typical lexical baselines. Results show that both the probabilistic model and our percentage-coverage baseline perform favorably relative to prior art. These results support the viability of the probabilistic framework while pointing at certain modeling aspects that need to be improved.

2 Probabilistic Model

Under the lexical entailment scope, our modeling goal is obtaining a probabilistic score for the likelihood that all H's terms are entailed by T. To that end, we model prominent aspects of lexical entailment, which were mostly neglected by previous lexical methods: (1) distinguishing different reliability levels of lexical resources; (2) allowing transitive chains of rule applications and considering their length when estimating their validity; and (3) considering multiple entailments when entailing a term.

[2]See ablation tests reports in http://aclweb.org/aclwiki/ index.php?title=RTE_Knowledge_Resources#Ablation_Tests

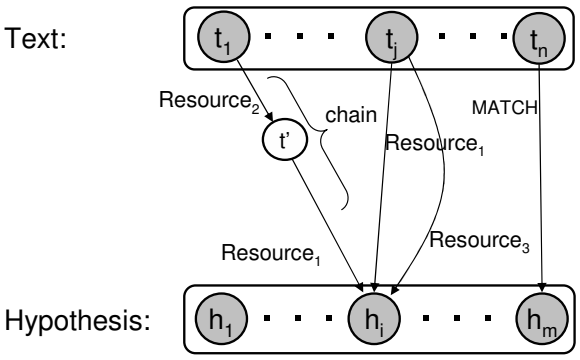

Figure 1: The generative process of entailing terms of a hypothesis from a text. Edges represent entailment rules. There are 3 evidences for the entailment of h_i: a rule from $Resource_1$, another one from $Resource_3$ both suggesting that t_j entails it, and a chain from t_1 through an intermediate term t'.

2.1 Model Description

For T to entail H it is usually a necessary, but not sufficient, that every term $h \in H$ would be entailed by at least one term $t \in T$ (Glickman et al., 2006). Figure 1 describes the process of entailing hypothesis terms. The trivial case is when identical terms, possibly at the stem or lemma level, appear in T and H (a direct match as t_n and h_m in Figure 1). Alternatively, we can establish entailment based on knowledge of entailing lexical-semantic relations, such as synonyms, hypernyms and morphological derivations, available in lexical resources (e.g the rule *inference* → *reasoning* from WordNet). We denote by $R(r)$ the resource which provided the rule r.

Since entailment is a transitive relation, rules may compose transitive *chain*s that connect a term $t \in T$ to a term $h \in H$ through intermediate terms. For instance, from the rules *infer* → *inference* and *inference* → *reasoning* we can deduce the rule *infer* → *reasoning* (were *inference* is the intermediate term as t' in Figure 1).

Multiple chains may connect t to h (as for t_j and h_i in Figure 1) or connect several terms in T to h (as t_1 and t_j are indicating the entailment of h_i in Figure 1), thus providing multiple evidence for h's entailment. It is reasonable to expect that if a term t indeed entails a term h, it is likely to find evidences for this relation in several resources.

Taking a probabilistic perspective, we assume a

parameter θ_R for each resource R, denoting its reliability, i.e. the prior probability that applying a rule from R corresponds to a valid entailment instance. Direct matches are considered as a special "resource", called MATCH, for which θ_{MATCH} is expected to be close to 1.

We now present our probabilistic model. For a text term $t \in T$ to entail a hypothesis term h by a chain c, denoted by $t \xrightarrow{c} h$, the application of every $r \in c$ must be valid. Note that a rule r in a chain c connects two terms (its left-hand-side and its right-hand-side, denoted *lhs* $\rightarrow$ *rhs*). The lhs of the first rule in c is $t \in T$ and the rhs of the last rule in it is $h \in H$. We denote the event of a valid rule application by $lhs \xrightarrow{r} rhs$. Since a-priori a rule r is valid with probability $\theta_{R(r)}$, and assuming independence of all $r \in c$, we obtain Eq. 1 to specify the probability of the event $t \xrightarrow{c} h$. Next, let $C(h)$ denote the set of chains which suggest the entailment of h. The probability that T does not entail h at all (by any chain), specified in Eq. 2, is the probability that all these chains are not valid. Finally, the probability that T entails all of H, assuming independence of H's terms, is the probability that every $h \in H$ is entailed, as given in Eq. 3. Notice that there could be a term h which is not covered by any available rule chain. Under this formulation, we assume that each such h is covered by a single rule coming from a special "resource" called UNCOVERED (expecting $\theta_{\text{UNCOVERED}}$ to be relatively small).

$$p(t \xrightarrow{c} h) = \prod_{r \in c} p(lhs \xrightarrow{r} rhs) = \prod_{r \in c} \theta_{R(r)} \quad (1)$$

$$p(T \nrightarrow h) = \prod_{c \in C(h)} [1 - p(t \xrightarrow{c} h)] \quad (2)$$

$$p(T \rightarrow H) = \prod_{h \in H} p(T \rightarrow h) \quad (3)$$

As can be seen, our model indeed distinguishes varying resource reliability, decreases entailment probability as rule chains grow and increases it when entailment of a term is supported by multiple chains.

The above treatment of uncovered terms in H, as captured in Eq. 3, assumes that their entailment probability is independent of the rest of the hypothesis. However, when the number of covered hypothesis terms increases the probability that the remaining terms are actually entailed by T increases too (even though we do not have supporting knowledge for their entailment). Thus, an alternative model is to group all uncovered terms together and estimate the overall probability of their joint entailment as a function of the lexical coverage of the hypothesis. We denote H_c as the subset of H's terms which are covered by some rule chain and H_{uc} as the remaining uncovered part. Eq. 3a then provides a refined entailment model for H, in which the second term specifies the probability that H_{uc} is entailed given that H_c is validly entailed and the corresponding lengths:

$$p(T \rightarrow H) = [\prod_{h \in H_c} p(T \rightarrow h)] \cdot p(T \rightarrow H_{uc} \mid |H_c|, |H|) \quad (3a)$$

2.2 Parameter Estimation

The difficulty in estimating the θ_R values is that these are term-level parameters while the RTE-training entailment annotation is given for the sentence-level. Therefore, we use EM-based estimation for the hidden parameters (Dempster et al., 1977). In the E step we use the current θ_R values to compute all $w_{hcr}(T, H)$ values for each training pair. $w_{hcr}(T, H)$ stands for the posterior probability that application of the rule r in the chain c for $h \in H$ is valid, given that either T entails H or not according to the training annotation (see Eq. 4). Remember that a rule r provides an entailment relation between its left-hand-side (*lhs*) and its right-hand-side (*rhs*). Therefore Eq. 4 uses the notation $lhs \xrightarrow{r} rhs$ to designate the application of the rule r (similar to Eq. 1).

$$E: \quad w_{hcr}(T, H) = \begin{cases} p(lhs \xrightarrow{r} rhs | T \rightarrow H) = \\ \frac{p(T \rightarrow H | lhs \xrightarrow{r} rhs) p(lhs \xrightarrow{r} rhs)}{p(T \rightarrow H)} & \text{if } T \rightarrow H \\ p(lhs \xrightarrow{r} rhs | T \nrightarrow H) = \\ \frac{p(T \nrightarrow H | lhs \xrightarrow{r} rhs) p(lhs \xrightarrow{r} rhs)}{p(T \nrightarrow H)} & \text{if } T \nrightarrow H \end{cases} \quad (4)$$

After applying Bayes' rule we get a fraction with Eq. 3 in its denominator and $\theta_{R(r)}$ as the second term of the numerator. The first numerator term is defined as in Eq. 3 except that for the corresponding rule application we substitute $\theta_{R(r)}$ by 1 (per the conditioning event). The probabilistic model defined by Eq. 1-3 is a loop-free directed acyclic graphical model

(aka a Bayesian network). Hence the E-step probabilities can be efficiently calculated using the belief propagation algorithm (Pearl, 1988).

The M step uses Eq. 5 to update the parameter set. For each resource R we average the $w_{hcr}(T, H)$ values for all its rule applications in the training, whose total number is denoted n_R.

$$M : \theta_R = \frac{1}{n_R} \sum_{T,H} \sum_{h \in H} \sum_{c \in C(h)} \sum_{r \in c | R(r)=R} w_{hcr}(T, H) \quad (5)$$

For Eq. 3a we need to estimate also $p(T \rightarrow H_{uc} \mid |H_c|, |H|)$. This is done directly via maximum likelihood estimation over the training set, by calculating the proportion of entailing examples within the set of all examples of a given hypothesis length ($|H|$) and a given number of covered terms ($|H_c|$). As $|H_c|$ we take the number of identical terms in T and H (exact match) since in almost all cases terms in H which have an exact match in T are indeed entailed. We also tried initializing the EM algorithm with these direct estimations but did not obtain performance improvements.

3 Evaluations and Results

The 5^{th} Recognizing Textual Entailment challenge (RTE-5) introduced a new search task (Bentivogli et al., 2009) which became the main task in RTE-6 (Bentivogli et al., 2010). In this task participants should find all sentences that entail a given hypothesis in a given document cluster. This task's data sets reflect a natural distribution of entailments in a corpus and demonstrate a more realistic scenario than the previous RTE challenges.

In our system, sentences are tokenized and stripped of stop words and terms are lemmatized and tagged for part-of-speech. As lexical resources we use WordNet (*WN*) (Fellbaum, 1998), taking as entailment rules synonyms, derivations, hyponyms and meronyms of the first senses of T and H terms, and the *CatVar* (Categorial Variation) database (Habash and Dorr, 2003). We allow rule chains of length up to 4 in WordNet (*WN*^{4}).

We compare our model to two types of baselines: (1) *RTE* published results: the average of the best runs of all systems, the best and second best performing lexical systems and the best full system of each challenge; (2) our implementation of lexical *coverage* model, tuning the percentage-of-coverage threshold for entailment on the training set. This model uses the same configuration as our *probabilistic* model. We also implemented an Information Retrieval style baseline[3] (both with and without lexical expansions), but given its poorer performance we omit its results here.

Table 1 presents the results. We can see that both our implemented models (probabilistic and coverage) outperform all RTE lexical baselines on both data sets, apart from (Majumdar and Bhattacharyya, 2010) which incorporates additional lexical resources, a named entity recognizer and a co-reference system. On RTE-5, the probabilistic model is comparable in performance to the best full system, while the coverage model achieves considerably better results. We notice that our implemented models successfully utilize resources to increase performance, as opposed to typical smaller or less consistent improvements in prior works (see Section 1).

	Model	F_1% RTE-5	F_1% RTE-6
RTE	avg. of all systems	30.5	33.8
	2^{nd} best lexical system	40.3[1]	44.0[2]
	best lexical system	44.4[3]	47.6[4]
	best full system	45.6[3]	48.0[5]
coverage	no resource	39.5	44.8
	+ WN	45.8	45.1
	+ CatVar	47.2	45.5
	+ WN + CatVar	48.5	44.7
	+ WN^4	46.3	43.1
probabilistic	no resource	41.8	42.1
	+ WN	45.0	45.3
	+ CatVar	42.0	45.9
	+ WN + CatVar	42.8	45.5
	+ WN^4	45.8	42.6

Table 1: Evaluation results on RTE-5 and RTE-6. RTE systems are: (1)(MacKinlay and Baldwin, 2009), (2)(Clark and Harrison, 2010), (3)(Mirkin et al., 2009)(2 submitted runs), (4)(Majumdar and Bhattacharyya, 2010) and (5)(Jia et al., 2010).

While the probabilistic and coverage models are comparable on RTE-6 (with non-significant advantage for the former), on RTE-5 the latter performs

[3]Utilizing Lucene search engine (http://lucene.apache.org)

better, suggesting that the probabilistic model needs to be further improved. In particular, WN^4 performs better than the single-step *WN* only on RTE-5, suggesting the need to improve the modeling of chaining. The fluctuations over the data sets and impacts of resources suggest the need for further investigation over additional data sets and resources. As for the coverage model, under our configuration it poses a bigger challenge for RTE systems than perviously reported baselines. It is thus proposed as an easy to implement baseline for future entailment research.

4 Conclusions and Future Work

This paper presented, for the first time, a principled and relatively rich probabilistic model for lexical entailment, amenable for estimation of hidden lexical-level parameters from standard sentence-level annotations. The positive results of the probabilistic model compared to prior art and its ability to exploit lexical resources indicate its future potential. Yet, further investigation is needed. For example, analyzing current model's limitations, we observed that the multiplicative nature of eqs. 1 and 3 (reflecting independence assumptions) is too restrictive, resembling a logical AND. Accordingly we plan to explore relaxing this strict conjunctive behavior through models such as noisy-AND (Pearl, 1988). We also intend to explore the contribution of our model, and particularly its estimated parameter values, within a complex system that integrates multiple levels of inference.

Acknowledgments

This work was partially supported by the NEGEV Consortium of the Israeli Ministry of Industry, Trade and Labor (www.negev-initiative.org), the PASCAL-2 Network of Excellence of the European Community FP7-ICT-2007-1-216886, the FIRB-Israel research project N. RBIN045PXH and by the Israel Science Foundation grant 1112/08.

References

Roy Bar-Haim, Jonathan Berant, Ido Dagan, Iddo Greental, Shachar Mirkin, Eyal Shnarch, and Idan Szpektor. 2008. Efficient semantic deduction and approximate matching over compact parse forests. In *Proceedings of Text Analysis Conference (TAC)*.

Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danilo Giampiccolo, and Bernardo Magnini. 2009. The fifth PASCAL recognizing textual entailment challenge. In *Proceedings of Text Analysis Conference (TAC)*.

Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. 2010. The sixth PASCAL recognizing textual entailment challenge. In *Proceedings of Text Analysis Conference (TAC)*.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

Aljoscha Burchardt, Nils Reiter, Stefan Thater, and Anette Frank. 2007. A semantic approach to textual entailment: System evaluation and task analysis. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Peter Clark and Phil Harrison. 2010. BLUE-Lite: a knowledge-based lexical entailment system for RTE6. In *Proceedings of Text Analysis Conference (TAC)*.

Courtney Corley and Rada Mihalcea. 2005. Measuring the semantic similarity of texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Lecture Notes in Computer Science*, volume 3944, pages 177–190.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society, series [B]*, 39(1):1–38.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Oren Glickman, Eyal Shnarch, and Ido Dagan. 2006. Lexical reference: a semantic matching subtask. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 172–179. Association for Computational Linguistics.

Nizar Habash and Bonnie Dorr. 2003. A categorial variation database for english. In *Proceedings of the North American Association for Computational Linguistics*.

Houping Jia, Xiaojiang Huang, Tengfei Ma, Xiaojun Wan, and Jianguo Xiao. 2010. PKUTM participation at TAC 2010 RTE and summarization track. In *Proceedings of Text Analysis Conference (TAC)*.

Bill MacCartney and Christopher D. Manning. 2007. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Andrew MacKinlay and Timothy Baldwin. 2009. A baseline approach to the RTE5 search pilot. In *Proceedings of Text Analysis Conference (TAC)*.

Debarghya Majumdar and Pushpak Bhattacharyya. 2010. Lexical based text entailment system for main task of RTE6. In *Proceedings of Text Analysis Conference (TAC)*.

Shachar Mirkin, Roy Bar-Haim, Jonathan Berant, Ido Dagan, Eyal Shnarch, Asher Stern, and Idan Szpektor. 2009. Addressing discourse and document structure in the RTE search task. In *Proceedings of Text Analysis Conference (TAC)*.

Judea Pearl. 1988. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.

Marta Tatu and Dan Moldovan. 2007. COGEX at RTE 3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*.

Rui Wang, Yi Zhang, and Guenter Neumann. 2009. A joint syntactic-semantic representation for recognizing textual relatedness. In *Proceedings of Text Analysis Conference (TAC)*.

Fabio Massimo Zanzotto and Alessandro Moschitti. 2006. Automatic learning of textual entailments with cross-pair similarities. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.