A Two Level Model for Context Sensitive Inference Rules

Oren Melamud[§], **Jonathan Berant**[†], **Ido Dagan**[§], **Jacob Goldberger**[♢], **Idan Szpektor**[‡]
§ Computer Science Department, Bar-Ilan University
† Computer Science Department, Stanford University
♢ Faculty of Engineering, Bar-Ilan University
‡ Yahoo! Research Israel
{melamuo,dagan,goldbej}@{cs,cs,eng}.biu.ac.il
joberant@stanford.edu
idan@yahoo-inc.com

Abstract

Automatic acquisition of inference rules for predicates has been commonly addressed by computing distributional similarity between vectors of argument words, operating at the word space level. A recent line of work, which addresses context sensitivity of rules, represented contexts in a latent topic space and computed similarity over topic vectors. We propose a novel two-level model, which computes similarities between word-level vectors that are biased by topic-level context representations. Evaluations on a naturally-distributed dataset show that our model significantly outperforms prior word-level and topic-level models. We also release a first context-sensitive inference rule set.

1 Introduction

Inference rules for predicates have been identified as an important component in semantic applications, such as Question Answering (QA) (Ravichandran and Hovy, 2002) and Information Extraction (IE) (Shinyama and Sekine, 2006). For example, the inference rule '*X treat Y* $\rightarrow$ *X relieve Y*' can be useful to extract pairs of drugs and the illnesses which they relieve, or to answer a question like "*Which drugs relieve headache?*". Along this vein, such inference rules constitute a crucial component in generic modeling of textual inference, under the Textual Entailment paradigm (Dagan et al., 2006; Dinu and Wang, 2009).

Motivated by these needs, substantial research was devoted to automatic learning of inference rules from corpora, mostly in an unsupervised distributional setting. This research line was mainly initiated by the highly-cited DIRT algorithm (Lin and Pantel, 2001), which learns inference for binary predicates with two argument slots (like the rule in the example above). DIRT represents a predicate by two vectors, one for each of the argument slots, where the vector entries correspond to the argument words that occurred with the predicate in the corpus. Inference rules between pairs of predicates are then identified by measuring the similarity between their corresponding argument vectors. This general scheme was further enhanced in several directions, e.g. directional similarity (Bhagat et al., 2007; Szpektor and Dagan, 2008) and meta-classification over similarity values (Berant et al., 2011). Consequently, several knowledge resources of inference rules were released, containing the top scoring rules for each predicate (Schoenmackers et al., 2010; Berant et al., 2011; Nakashole et al., 2012).

The above mentioned methods provide a single confidence score for each rule, which is based on the obtained degree of argument-vector similarities. Thus, a system that applies an inference rule to a text may estimate the validity of the rule application based on the pre-specified rule score. However, the validity of an inference rule may depend on the context in which it is applied, such as the context specified by the given predicate's arguments. For example, '*AT&T acquire T-Mobile* $\rightarrow$ *AT&T purchase T-Mobile*', is a valid application of the rule '*X acquire Y* $\rightarrow$ *X purchase Y*', while '*Children acquire skills* $\rightarrow$ *Children purchase skills*' is not. To address this issue, a line of works emerged which computes a *context-sensitive* reliability score for each rule *application*, based on the given context.

The major trend in context-sensitive inference models utilizes latent or class-based methods for context modeling (Pantel et al., 2007; Szpektor et al., 2008; Ritter et al., 2010; Dinu and Lapata, 2010b). In particular, the more recent methods (Ritter et al., 2010; Dinu and Lapata, 2010b) modeled predicates in context as a probability distribution over topics learned by a Latent Dirichlet Allo-

cation (LDA) model. Then, similarity is measured between the two topic distribution vectors corresponding to the two sides of the rule in the given context, yielding a context-sensitive score for each particular rule application.

We notice at this point that while context-insensitive methods represent predicates by argument vectors in the original fine-grained word space, context-sensitive methods represent them as vectors at the level of latent topics. This raises the question of whether such coarse-grained topic vectors might be less informative in determining the semantic similarity between the two predicates.

To address this hypothesized caveat of prior context-sensitive rule scoring methods, we propose a novel generic scheme that integrates word-level and topic-level representations. Our scheme can be applied on top of any context-insensitive "base" similarity measure for rule learning, which operates at the word level, such as Cosine or Lin (Lin, 1998). Rather than computing a single context-insensitive rule score, we compute a distinct word-level similarity score for each topic in an LDA model. Then, when applying a rule in a given context, these different scores are weighed together based on the specific topic distribution under the given context. This way, we calculate similarity over vectors in the original word space, while biasing them towards the given context via a topic model.

In order to promote replicability and equal-term comparison with our results, we based our experiments on publicly available datasets, both for unsupervised learning of the evaluated models and for testing them over a random sample of rule applications. We apply our two-level scheme over three state-of-the-art context-insensitive similarity measures. The evaluation compares performances both with the original context-insensitive measures and with recent LDA-based context-sensitive methods, showing consistent and robust advantages of our scheme. Finally, we release a context-sensitive rule resource comprising over 2,000 frequent verbs and one million rules.

2 Background and Model Setting

This section presents components of prior work which are included in our model and experiments, setting the technical preliminaries for the rest of the paper. We first present context-insensitive rule learning, based on distributional similarity at the word level, and then context-sensitive scoring for rule applications, based on topic-level similarity. Some further discussion of related work appears in Section 6.

2.1 Context-insensitive Rule Learning

A predicate inference rule '*LHS* $\rightarrow$ *RHS*', such as 'X *acquire* Y $\rightarrow$ X *purchase* Y', specifies a directional inference relation between two predicates. Each rule side consists of a lexical predicate and (two) variable slots for its arguments.[1] Different representations have been used to specify predicates and their argument slots, such as word lemma sequences, regular expressions and dependency parse fragments. A rule can be *applied* when its LHS matches a predicate with a pair of arguments in a text, allowing us to infer its RHS, with the corresponding instantiations for the argument variables. For example, given the text "*AT&T acquires T-Mobile*", the above rule infers "*AT&T purchases T-Mobile*".

The DIRT algorithm (Lin and Pantel, 2001) follows the distributional similarity paradigm to learn predicate inference rules. For each predicate, DIRT represents each of its argument slots by an *argument vector*. We denote the two vectors of the X and Y slots of a predicate *pred* by v^x_{pred} and v^y_{pred}, respectively. Each entry of a vector v corresponds to a particular word (or term) w that instantiated the argument slot in a learning corpus, with a value $v(w) = PMI(pred, w)$ (with PMI standing for point-wise mutual information).

To learn inference rules, DIRT considers (in principle) each pair of binary predicates that occurred in the corpus for a candidate rule, '*LHS* $\rightarrow$ *RHS*'. Then, DIRT computes a *reliability score* for the rule by combining the measured similarities between the corresponding argument vectors of the two rule sides. Concretely, denoting by l and r the predicates appearing in the two rule sides, DIRT's reliability score is defined as follows:

$$\begin{aligned}&\text{score}_{\text{DIRT}}(\textit{LHS} \rightarrow \textit{RHS}) \\ &= \sqrt{\text{sim}(v^x_l, v^x_r) \cdot \text{sim}(v^y_l, v^y_r)}\end{aligned} \quad (1)$$

where $\text{sim}(v, v')$ is a vector similarity measure. Specifically, DIRT employs the Lin similarity

[1] We follow most of the inference-rule learning literature, which focused on binary predicates. However, our context-sensitive scheme can be applied to any arity.

measure from (Lin, 1998), defined as follows:

$$Lin(v, v') = \frac{\sum_{w \in v \cap v'}[v(w) + v'(w)]}{\sum_{w \in v \cup v'}[v(w) + v'(w)]} \quad (2)$$

We note that the general DIRT scheme may be used while employing other "base" vector similarity measures. For example, the Lin measure is symmetric, and thus using it would yield the same reliability score when swapping the two sides of a rule. This issue has been addressed in a separate line of research which introduced directional similarity measures suitable for inference relations (Bhagat et al., 2007; Szpektor and Dagan, 2008; Kotlerman et al., 2010). In our experiments we apply our proposed context-sensitive similarity scheme over three different base similarity measures.

DIRT and similar context-insensitive inference methods provide a single reliability score for a learned inference rule, which aims to predict the validity of the rule's applications. However, as exemplified in the Introduction, an inference rule may be valid in some contexts but invalid in others (e.g. *acquiring* entails *purchasing* for *goods*, but not for *skills*). Since vector similarity in DIRT is computed over the single aggregate argument vector, the obtained reliability score tends to be biased towards the dominant contexts of the involved predicates. For example, we may expect a higher score for '*acquire* → *purchase*' than for '*acquire* → *learn*', since the former matches a more frequent sense of *acquire* in a typical corpus. Following this observation, it is desired to obtain a *context-sensitive reliability score* for each rule application in a given context, as described next.

2.2 Context-sensitive Rule Applications

To assess the reliability of applying an inference rule in a given context we need some model for context representation, that should affect the rule reliability score. A major trend in past work is to represent contexts in a reduced-dimensionality latent or class-based model. A couple of earlier works utilized a cluster-based model (Pantel et al., 2007) and an LSA-based model (Szpektor et al., 2008), in a selectional-preferences style approach. Several more recent works utilize a Latent Dirichlet Allocation (LDA) (Blei et al., 2003) framework. We now present an underlying unified view of the *topic-level* models in (Ritter et al., 2010; Dinu and Lapata, 2010b), which we follow in our own model and in comparative model evaluations. We note that a similar LDA model construction was employed also in (Séaghdha, 2010), for estimating predicate-argument likelihood.

First, an LDA model is constructed, as follows. Similar to the construction of argument vectors in the distributional model (described above in subsection 2.1), all arguments instantiating each predicate slot are extracted from a large learning corpus. Then, for each slot of each predicate, a pseudo-document is constructed containing the set of all argument words that instantiated this slot in the corpus. We denote the two documents constructed for the X and Y slots of a predicate *pred* by d^x_{pred} and d^y_{pred}, respectively. In comparison to the distributional model, these two documents correspond to the analogous argument vectors v^x_{pred} and v^y_{pred}, both containing exactly the same set of words.

Next, an LDA model is learned from the set of all pseudo-documents, extracted for all predicates.[2] The learning process results in the construction of K latent topics, where each topic t specifies a distribution over all words, denoted by $p(w|t)$, and a topic distribution for each pseudo-document d, denoted by $p(t|d)$.

Within the LDA model we can derive the a-posteriori topic distribution conditioned on a particular word within a document, denoted by $p(t|d, w) \propto p(w|t) \cdot p(t|d)$. In the topic-level model, d corresponds to a predicate slot and w to a particular argument word instantiating this slot. Hence, $p(t|d, w)$ is viewed as specifying the relevance (or likelihood) of the topic t for the predicate slot in the context of the given argument instantiation. For example, for the predicate slot 'acquire *Y*' in the context of the argument *'IBM'*, we expect high relevance for a topic about companies, while in the context of the argument *'knowledge'* we expect high relevance for a topic about abstract concepts. Accordingly, the distribution $p(t|d, w)$ over all topics provides a topic-level representation for a predicate slot in the context of a particular argument w. This representation is used by the topic-level model to compute a context-sensitive score for inference rule applications, as follows.

[2] We note that there are variants in the type of LDA model and the way the pseudo-documents are constructed in the referenced prior work. In order to focus on the inference methods rather than on the underlying LDA model, we use the LDA framework described in this paper for all compared methods.

Consider the application of an inference rule '*LHS* $\rightarrow$ *RHS*' in the context of a particular pair of arguments for the X and Y slots, denoted by w_x and w_y, respectively. Denoting by l and r the predicates appearing in the two rule sides, the reliability score of the topic-level model is defined as follows (we present a geometric mean formulation for consistency with DIRT):

$$\begin{aligned}&\text{score}_{\text{Topic}}(\textit{LHS} \rightarrow \textit{RHS}, w_x, w_y) \\ &= \sqrt{\text{sim}(d_l^x, d_r^x, w_x) \cdot \text{sim}(d_l^y, d_r^y, w_y)}\end{aligned} \quad (3)$$

where $\text{sim}(d, d', w)$ is a topic-distribution similarity measure conditioned on a given context word. Specifically, Ritter et al. (2010) utilized the dot product form for their similarity measure:

$$\text{sim}_{\text{DC}}(d, d', w) = \Sigma_t[p(t|d, w) \cdot p(t|d', w)] \quad (4)$$

(the subscript DC stands for double-conditioning, as both distributions are conditioned on the argument word, unlike the measure below).

Dinu and Lapata (2010b) presented a slightly different similarity measure for topic distributions that performed better in their setting as well as in a related later paper on context-sensitive scoring of lexical similarity (Dinu and Lapata, 2010a). In this measure, the topic distribution for the right hand side of the rule is not conditioned on w:

$$\text{sim}_{\text{SC}}(d, d', w) = \Sigma_t[p(t|d, w) \cdot p(t|d')] \quad (5)$$

(the subscript SC stands for single-conditioning, as only the left distribution is conditioned on the argument word). They also experimented with a few variants for the structure of the similarity measure and assessed that best results are obtained with the dot product form. In our experiments, we employ these two similarity measures for topic distributions as baselines representing topic-level models.

Comparing the context-insensitive and context-sensitive models, we see that both of them measure similarity between vector representations of corresponding predicate slots. However, while DIRT computes $\text{sim}(v, v')$ over vectors in the original word-level space, topic-level models compute $\text{sim}(d, d', w)$ by measuring similarity of vectors in a reduced-dimensionality latent space. As conjectured in the introduction, such coarse-grain representation might lead to loss of information. Hence, in the next section we propose a combined two-level model, which represents predicate slots in the original word-level space while biasing the similarity measure through topic-level context models.

3 Two-level Context-sensitive Inference

Our model follows the general DIRT scheme while extending it to handle context-sensitive scoring of rule applications, addressing the scenario dealt by the context-sensitive topic models. In particular, we define the context-sensitive score score_{WT}, where WT stands for the combination of the Word/Topic levels:

$$\begin{aligned}&\text{score}_{\text{WT}}(\textit{LHS} \rightarrow \textit{RHS}, w_x, w_y) \\ &= \sqrt{\text{sim}(v_l^x, v_r^x, w_x) \cdot \text{sim}(v_l^y, v_r^y, w_y)}\end{aligned} \quad (6)$$

Thus, our model computes similarity over word-level (rather than topic-level) argument vectors, while biasing it according to the specific argument words in the given rule application context. The core of our contribution is thus defining the context-sensitive word-level vector similarity measure $\text{sim}(v, v', w)$, as described in the remainder of this section.

Following the methods in Section 2, for each predicate *pred* we construct, from the learning corpus, its argument vectors v_{pred}^x and v_{pred}^y as well as its argument pseudo-documents d_{pred}^x and d_{pred}^y. For convenience, when referring to an argument vector v, we will denote the corresponding pseudo-document by d_v. Based on all pseudo-documents we learn an LDA model and obtain its associated probability distributions.

The calculation of $\text{sim}(v, v', w)$ is composed of two steps. At learning time, we compute for each candidate rule a separate, topic-biased, similarity score per each of the topics in the LDA model. Then, at rule application time, we compute an overall reliability score for the rule by combining the per-topic similarity scores, while biasing the score combination according to the given context of w. These two steps are described in the following two subsections.

3.1 Topic-biased Word-vector Similarities

Given a pair of word vectors v and v', and any desired "base" vector similarity measure sim (*e.g.* sim_{Lin}), we compute a *topic-biased* similarity score for each LDA topic t, denoted by $\text{sim}_t(v, v')$. $\text{sim}_t(v, v')$ is computed by applying

the original similarity measure over topic-biased versions of v and v', denoted by v_t and v'_t:

$$\mathrm{sim}_t(v, v') = \mathrm{sim}(v_t, v'_t)$$

where

$$v_t(w) = v(w) \cdot p(t|d_v, w)$$

That is, each value in the biased vector, $v_t(w)$, is obtained by weighing the original value $v(w)$ by the relevance of the topic t to the argument word w within d_v. This way, rather than replacing altogether the word-level values $v(w)$ by the topic probabilities $p(t|d_v, w)$, as done in the topic-level models, we use the latter to only bias the former while preserving fine-grained word-level representations. The notation Lin_t denotes the sim_t measure when applied using Lin as the base similarity measure sim.

This learning process results in K different topic-biased similarity scores for each candidate rule, where K is the number of LDA topics. Table 1 illustrates topic-biased similarities for the Y slot of two rules involving the predicate '*acquire*'. As can be seen, the topic-biased score Lin_t for '*acquire* → *learn*' for t_2 is higher than the Lin score, since this topic is characterized by arguments that commonly appear with both predicates of the rule. Consequently, the two predicates are found to be distributionally similar when biased for this topic. On the other hand, the topic-biased similarity for t_1 is substantially lower, since prominent words in this topic are likely to occur with '*acquire*' but not with '*learn*', yielding low distributional similarity. Opposite behavior is exhibited for the rule '*acquire* → *purchase*'.

Topic	t_1	t_2
Top 5 words	calbiochem corel networks viacom financially	rights syndrome majority knowledge skill
acquire → *learn*		
$\mathrm{Lin}_t(v, v')$	0.040	0.334
$\mathrm{Lin}(v, v')$	0.165	
acquire → *purchase*		
$\mathrm{Lin}_t(v, v')$	0.427	0.241
$\mathrm{Lin}(v, v')$	0.267	

Table 1: Two characteristic topics for the Y slot of '*acquire*', along with their topic-biased Lin similarities scores Lin_t, compared with the original Lin similarity, for two rules. The relevance of each topic to different arguments of '*acquire*' is illustrated by showing the top 5 words in the argument vector $v^y_{acquire}$ for which the illustrated topic is the most likely one.

3.2 Context-sensitive Similarity

When applying an inference rule, we compute for each slot its context-sensitive similarity score $\mathrm{sim}_{\mathrm{WT}}(v, v', w)$, where v and v' are the slot's argument vectors for the two rule sides and w is the word instantiating the slot in the given rule application. This score is computed as a weighted average of the rule's K topic-biased similarity scores sim_t. In this average, each topic is weighed by its "relevance" for the context in which the rule is applied, which consists of the left-hand-side predicate v and the argument w. This relevance is captured by $p(t|d_v, w)$:

$$\mathrm{sim}_{\mathrm{WT}}(v, v', w) = \sum_t [p(t|d_v, w) \cdot \mathrm{sim}_t(v, v')] \quad (7)$$

This way, a rule application would obtain a high score only if the current context fits those topics for which the rule is indeed likely to be valid, as captured by a high topic-biased similarity. The notation Lin_{WT} denotes the $\mathrm{sim}_{\mathrm{WT}}$ measure, when using Lin_t as the topic-biased similarity measure.

Table 2 illustrates the calculation of context-sensitive similarity scores in four rule applications, involving the Y slot of the predicate '*acquire*'. We observe that relative to the fixed context-insensitive Lin score, the score of '*acquire* → *learn*' is substantially promoted for the argument '*skill*' while being demoted for '*Skype*'. The opposite behavior is observed for '*acquire* → *purchase*', altogether demonstrating how our model successfully biases the similarity score according to rule validity in context.

4 Experimental Settings

To evaluate our model, we compare it both to context-insensitive similarity measures as well as to prior context-sensitive methods. Furthermore, to better understand its applicability in typical NLP tasks, we focus on an evaluation setting that corresponds to a natural distribution of examples from a large corpus.

Topic	t_1	t_2
Top 5 words	calbiochem corel networks viacom financially	rights syndrome majority knowledge skill
'*acquire Skype → learn Skype*'		
$p(t\|d_v, w)$	0.974	0.000
$\text{Lin}_t(v, v')$	0.040	0.334
$\text{Lin}_{\text{WT}}(v, v', w)$	**0.039**	
$\text{Lin}(v, v')$	0.165	
'*acquire Skype → purchase Skype*'		
$p(t\|d_v, w)$	0.974	0.000
$\text{Lin}_t(v, v')$	0.427	0.241
$\text{Lin}_{\text{WT}}(v, v', w)$	**0.417**	
$\text{Lin}(v, v')$	0.267	
'*acquire skill → learn skill*'		
$p(t\|d_v, w)$	0.000	0.380
$\text{Lin}_t(v, v')$	0.040	0.334
$\text{Lin}_{\text{WT}}(v, v', w)$	**0.251**	
$\text{Lin}(v, v')$	0.165	
'*acquire skill → purchase skill*'		
$p(t\|d_v, w)$	0.000	0.380
$\text{Lin}_t(v, v')$	0.427	0.241
$\text{Lin}_{\text{WT}}(v, v', w)$	**0.181**	
$\text{Lin}(v, v')$	0.267	

Table 2: Context-sensitive similarity scores (in bold) for the Y slots of four rule applications. The components of the score calculation are shown for the topics of Table 1. For each rule application, the table shows a couple of the topic-biased scores Lin_t of the rule (as in Table 1), along with the topic relevance for the given context $p(t|d_v, w)$, which weighs the topic-biased scores in the Lin_{WT} calculation. The context-insensitive Lin score is shown for comparison.

4.1 Evaluated Rule Application Methods

We evaluated the following rule application methods: the original context-insensitive word model, following DIRT (Lin and Pantel, 2001), as described in Equation 1, denoted by CI; our own topic-word context-sensitive model, as described in Equation 6, denoted by WT. In addition, we evaluated two variants of the topic-level context-sensitive model, denoted DC and SC. DC follows the double conditioned contextualized similarity measure according to Equation 4, as implemented by (Ritter et al., 2010), while SC follows the single conditioned one at Equation 5, as implemented by (Dinu and Lapata, 2010b; Dinu and Lapata, 2010a).

Since our model can contextualize various distributional similarity measures, we evaluated the performance of all the above methods on several base similarity measures and their learned rule-sets, namely Lin (Lin, 1998), BInc (Szpektor and Dagan, 2008) and vector Cosine similarity. The Lin similarity measure is described in Equation 2. Binc (Szpektor and Dagan, 2008) is a directional similarity measure between word vectors, which outperformed Lin for predicate inference (Szpektor and Dagan, 2008).

To build the rule-sets and models for the tested approaches we utilized the ReVerb corpus (Fader et al., 2011), a large scale publicly available web-based open extractions data set, containing about 15 million unique template extractions.[3] ReVerb template extractions/instantiations are in the form of a tuple $(x, pred, y)$, containing $pred$, a verb predicate, x, the argument instantiation of the template's slot X, and y, the instantiation of the template's slot Y.

ReVerb includes over 600,000 different templates that comprise a verb but may also include other words, for example '*X* can accommodate up to *Y*'. Yet, many of these templates share a similar meaning, *e.g.* '*X* accommodate up to *Y*', '*X* can accommodate up to *Y*', '*X* will accommodate up to *Y*', etc. Following Sekine (2005), we clustered templates that share their main verb predicate in order to scale down the number of different predicates in the corpus and collect richer word co-occurrence statistics per predicate.

Next, we applied some clean-up preprocessing to the ReVerb extractions. This includes discarding stop words, rare words and non-alphabetical words instantiating either the X or the Y arguments. In addition, we discarded all predicates that co-occur with less than 100 unique argument words in each slot. The remaining corpus consists of 7 million unique extractions and 2,155 verb predicates.

Finally, we trained an LDA model, as described in Section 2, using Mallet (McCallum, 2002). Then, for each original context-insensitive similarity measure, we learned from ReVerb a rule-set comprised of the top 500 rules for every identified predicate. To complete the learning, we calculated the topic-biased similarity score for each learned rule under each LDA topic, as specified in our context-sensitive model. We release a rule set comprising the top 500 context-sensitive rules that we learned for each of the verb predicates in our learning corpus, along with our trained LDA

[3] ReVerb is available at `http://reverb.cs.washington.edu/`

Method	Lin	BInc	Cosine
Valid	266	254	272
Invalid	545	523	539
Total	811	777	811

Table 3: Sizes of rule application test set for each learned rule-set.

model.[4]

4.2 Evaluation Task

To evaluate the performance of the different methods we chose the dataset constructed by Zeichner et al. (2012). [5] This publicly available dataset contains about 6,500 manually annotated predicate template rule applications, each one labeled as correct or incorrect. For example, '*Jack agree with Jill* $\nrightarrow$ *Jack feel sorry for Jill*' is a rule application in this dataset, labeled as incorrect, and '*Registration open this month* $\rightarrow$ *Registration begin this month*' is another rule application, labeled as correct. Rule applications were generated by randomly sampling extractions from ReVerb, such as *('Jack','agree with','Jill')* and then sampling possible rules for each, such as '*agree with* $\rightarrow$ *feel sorry for*'. Hence, this dataset provides naturally distributed rule inferences with respect to ReVerb.

Whenever we evaluated a distributional similarity measure (namely Lin, BInc, or Cosine), we discarded instances from Zeichner et al.'s dataset in which the assessed rule is not in the context-insensitive rule-set learned for this measure or the argument instantiation of the rule is not in the LDA lexicon. We refer to the remaining instances as the *test set* per measure, *e.g.* Lin's test set. Table 3 details the size of each such test set in our experiment.

Finally, the task under which we assessed the tested models is to rank all rule applications in each test set, aiming to rank the valid rule applications above the invalid ones.

5 Results

We evaluated the performance of each tested method by measuring Mean Average Precision (MAP) (Manning et al., 2008) of the rule application ranking computed by this method. In order to compute MAP values and corresponding statistical significance, we randomly split each test set into 30 subsets. For each method we computed Average Precision on every subset and then took the average over all subsets as the MAP value.

Method	Lin	BInc	Cosine
CI	0.503	0.513	0.513
DC	0.451 (1200)	0.455 (1200)	0.455 (1200)
SC	0.443 (1200)	0.458 (1200)	0.452 (1200)
WT	**0.562** (100)	**0.584** (50)	**0.565** (25)

Table 4: MAP values on corresponding test set obtained by each method. Figures in parentheses indicate optimal number of LDA topics.

Since all tested context-sensitive approaches are based on LDA topics, we varied for each method the number of LDA topics K that optimizes its performance, ranging from 25 to 1600 topics. We used LDA hyperparameters $\beta = 0.01$ and $\alpha = 0.1$ for $K < 600$ and $\alpha = \frac{50}{K}$ for $K >= 600$.

Table 4 presents the optimal MAP performance of each tested measure. Our main result is that our model outperforms all other methods, both context-insensitive and context-sensitive, by a relative increase of more than 10% for all three similarity measures that we tested. This improvement is statistically significant at $p < 0.01$ for BInc and Lin, and $p < 0.015$ for Cosine, using paired t-test. This shows that our model indeed successfully leverages contextual information beyond the basic context-agnostic rule scores and is robust across measures.

Surprisingly, both baseline topic-level context-sensitive methods, namely DC and SC, underperformed compared to their context-insensitive baselines. While Dinu and Lapata (Dinu and Lapata, 2010b) did show improvement over context-insensitive DIRT, this result was obtained on the verbs of the Lexical Substitution Task in SemEval (McCarthy and Navigli, 2007), which was manually created with a bias for context-sensitive substitutions. However, our result suggests that topic-level models might not be robust enough when applied to a random sample of inferences.

An interesting indication of the differences between our word-topic model, WT, and topic-only models, DC and SC, lies in the optimal number of LDA topics required for each method. The number of topics in the range 25-100 performed almost equally well under the WT model for all base measures, with a moderate decline for higher numbers.

[4] Our resource is available at: `http://www.cs.biu.ac.il/~nlp/downloads/wt-rules.html`

[5] The dataset is available at: `http://www.cs.biu.ac.il/~nlp/downloads/annotation-rule-application.htm`

The need for this rather small number of topics is due to the nature of utilization of topics in WT. Specifically, topics are leveraged for high-level domain disambiguation, while fine grained word-level distributional similarity is computed for each rule under each such domain. This works best for a relatively low number of topics. However, in higher numbers, topics relate to narrower domains and then topic biased word level similarity may become less effective due to potential sparseness. On the other hand, DC and SC rely on topics as a surrogate to predicate-argument co-occurrence features, and thus require a relatively large number of them to be effective.

Delving deeper into our test-set, Zeichner et al. provided a more detailed annotation for each invalid rule application. Specifically, they annotated whether the context under which the rule is applied is valid. For example, in '*John bought my car* $\nrightarrow$ *John sold my car*' the inference is invalid due to an inherently incorrect rule, but the context is valid. On the other hand in '*my boss raised my salary* $\nrightarrow$ *my boss constructed my salary*' the context {*'my boss'*, *'my salary'*} for applying '*raise* $\rightarrow$ *construct*' is invalid. Following, we split the test-set for the base Lin measure into two test-sets: (a) test-set$_{\text{vc}}$, which includes all correct rule applications and incorrect ones only under valid contexts, and (b) test-set$_{\text{ivc}}$, which includes again all correct rule applications but incorrect ones only under invalid contexts.

Table 5 presents the performance of each compared method on the two test sets. On test-set$_{\text{ivc}}$, where context mismatches are abundant, our model outperformed all other baselines (statistically significant at $p < 0.01$). In addition, this time DC slightly outperformed CI. This result more explicitly shows the advantages of integrating word-level and context-sensitive topic-level similarities for differentiating valid and invalid contexts for rule applications. Yet, many invalid rule applications occur under valid contexts due to inherently incorrect rules, and we want to make sure that also in this scenario our model does not fall behind the context-insensitive measure. Indeed, on test-set$_{\text{vc}}$, in which context mismatches are rare, our algorithm is still better than the original measure, indicating that WT can be safely applied to distributional similarity measures without concerns of reduced performance in different context scenarios.

	test-set$_{\text{ivc}}$	test-set$_{\text{vc}}$
Size (valid:invalid)	432 (266:166)	645 (266:379)
CI	0.780	0.587
DC	0.796	0.498
SC	0.779	0.512
WT	**0.854**	**0.621**

Table 5: MAP results for the two split Lin test-sets.

6 Discussion and Future Work

This paper addressed the problem of computing context-sensitive reliability scores for predicate inference rules. In particular, we proposed a novel scheme that applies over any base distributional similarity measure which operates at the word level, and computes a single context-insensitive score for a rule. Based on such a measure, our scheme constructs a context-sensitive similarity measure that computes a reliability score for predicate inference rules applications in the context of given arguments.

The contextualization of the base similarity score was obtained using a topic-level LDA model, which was used in a novel way. First, it provides a topic bias for learning separate per-topic word-level similarity scores between predicates. Then, given a specific candidate rule application, the LDA model is used to infer the topic distribution relevant to the context specified by the given arguments. Finally, the context-sensitive rule application score is computed as a weighted average of the per-topic word-level similarity scores, which are weighed according to the inferred topic distribution.

While most works on context-insensitive predicate inference rules, such as DIRT (Lin and Pantel, 2001), are based on word-level similarity measures, almost all prior models addressing context-sensitive predicate inference rules are based on topic models (except for (Pantel et al., 2007), which was outperformed by later models). We therefore focused on comparing the performance of our two-level scheme with state-of-the-art prior topic-level and word-level models of distributional similarity, over a random sample of inference rule applications. Under this natural setting, the two-level scheme consistently outperformed both types of models when tested with three different base similarity measures. Notably, our model shows stable performance over a large subset of the data

where context sensitivity is rare, while topic-level models tend to underperform in such cases compared to the base context-insensitive methods.

Our work is closely related to another research line that addresses lexical similarity and substitution scenarios in context. While we focus on lexical-syntactic predicate templates and instantiations of their argument slots as context, lexical similarity methods consider various lexical units that are not necessarily predicates, with their context typically being the collection of words in a window around them.

Various approaches have been proposed to address lexical similarity. A number of works are based on a compositional semantics approach, where a prior representation of a target lexical unit is composed with the representations of words in its given context (Mitchell and Lapata, 2008; Erk and Padó, 2008; Thater et al., 2010). Other works (Erk and Padó, 2010; Reisinger and Mooney, 2010) use a rather large word window around target words and compute similarities between clusters comprising instances of word windows. In addition, (Dinu and Lapata, 2010a) adapted the predicate inference topic model from (Dinu and Lapata, 2010b) to compute lexical similarity in context.

A natural extension of our work would be to extend our two level model to accommodate context-sensitive lexical similarity. For this purpose we will need to redefine the scope of context in our model, and adapt our method to compute context-biased lexical similarities accordingly. Then we will also be able to evaluate our model on the Lexical Substitution Task (McCarthy and Navigli, 2007), which has been commonly used in recent years as a benchmark for context-sensitive lexical similarity models.

In a different NLP task, Eidelman et al. (2012) utilize a similar approach to ours for improving the performance of statistical machine translation (SMT). They learn an LDA model on the source language side of the training corpus with the purpose of identifying implicit sub-domains. Then they utilize the distribution over topics inferred for each document in their corpus to compute separate per-topic translation probability tables. Finally, they train a classifier to translate a given target word based on these tables and the inferred topic distribution of the given document in which the target word appears. A notable difference between our approach and theirs is that we use predicate pseudo-documents consisting of argument instantiations to learn our LDA model, while Eidelman et al. use the real documents in a corpus. We believe that combining these two approaches may improve performance for both textual inference and SMT and plan to experiment with this direction in future work.

Acknowledgments

This work was partially supported by the Israeli Ministry of Science and Technology grant 3-8705, the Israel Science Foundation grant 880/12, and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 287923 (EXCITEMENT).

References

Jonathan Berant, Ido Dagan, and Jacob Goldberger. 2011. Global learning of typed entailment rules. In *ACL*.

Rahul Bhagat, Patrick Pantel, Eduard Hovy, and Marina Rey. 2007. Ledir: An unsupervised algorithm for learning directionality of inference rules. In *Proceedings of EMNLP-CoNLL*.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

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.

Georgiana Dinu and Mirella Lapata. 2010a. Measuring distributional similarity in context. In *Proceedings of EMNLP*.

Georgiana Dinu and Mirella Lapata. 2010b. Topic models for meaning similarity in context. In *Proceedings of COLING: Posters*.

Georgiana Dinu and Rui Wang. 2009. Inference rules and their application to recognizing textual entailment. In *Proceedings EACL*.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proceedings of the ACL conference short papers*.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP*.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the ACL conference short papers*.

Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying relations for open information extraction. In *Proceedings of EMNLP*.

Lili Kotlerman, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Dekang Lin and Patrick Pantel. 2001. DIRT – discovery of inference rules from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2001*.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL*.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of SemEval*.

Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proceedings of ACL-08: HLT*.

Ndapandula Nakashole, Gerhard Weikum, and Fabian Suchanek. 2012. Patty: A taxonomy of relational patterns with semantic types. *EMNLP12*.

Patrick Pantel, Rahul Bhagat, Bonaventura Coppola, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*.

Deepak Ravichandran and Eduard Hovy. 2002. Learning surface text patterns for a question answering system. In *Proceedings of ACL*.

Joseph Reisinger and Raymond J Mooney. 2010. Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Alan Ritter, Oren Etzioni, et al. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings of ACL*.

Stefan Schoenmackers, Jesse Davis, Oren Etzioni, and Daniel Weld. 2010. Learning first-order horn clauses from web text. In *Proceedings of EMNLP*.

Diarmuid O Séaghdha. 2010. Latent variable models of selectional preference. In *Proceedings of ACL*.

Satoshi Sekine. 2005. Automatic paraphrase discovery based on context and keywords between ne pairs. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Yusuke Shinyama and Satoshi Sekine. 2006. Preemptive information extraction using unrestricted relation discovery. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*.

Idan Szpektor and Ido Dagan. 2008. Learning entailment rules for unary templates. In *Proceedings of COLING*.

Idan Szpektor, Ido Dagan, Roy Bar-Haim, and Jacob Goldberger. 2008. Contextual preferences. In *Proceedings of ACL-08: HLT*.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2010. Contextualizing semantic representations using syntactically enriched vector models. In *Proceedings of ACL*.

Naomi Zeichner, Jonathan Berant, and Ido Dagan. 2012. Crowdsourcing inference-rule evaluation. In *Proceedings of ACL (short papers)*.