# A Latent Discriminative Model for Compositional Entailment Relation Recognition Using Natural Logic

*Yotaro Watanabe*[1] *Junta Mizuno*[1] *Eric Nichols*[1]
*Naoaki Okazaki*[1,2] *Kentaro Inui*[1]
(1) Graduate School of Information Sciences, Tohoku University
(2) Precursory Research for Embryonic Science and Technology, Japan Science and Technology Agency
{yotaro-w,junta-m,eric,okazaki,inui}@ecei.tohoku.ac.jp

ABSTRACT
Recognizing semantic relations between sentences, such as entailment and contradiction, is a challenging task that requires detailed analysis of the interaction between diverse linguistic phenomena. In this paper, we propose a latent discriminative model that unifies a statistical framework and a theory of Natural Logic to capture complex interactions between linguistic phenomena. The proposed approach jointly models alignments, their local semantic relations, and a sentence-level semantic relation, and has hidden variables including alignment edits between sentences and their semantic relations, only requires sentences pairs annotated with sentence-level semantic relations as training data to learn appropriate alignments. In evaluation on a dataset including diverse linguistic phenomena, our proposed method achieved a competitive results on alignment prediction, and significant improvements on a sentence-level semantic relation recognition task compared to an alignment supervised model. Our analysis did not provide evidence that directly learning alignments and their labels using gold standard alignments contributed to semantic relation recognition performance and instead suggests that they can be detrimental to performance if used in a manner that prevents the learning of globally optimal alignments.

KEYWORDS: Recognizing Textual Entailment, Natural Logic, Latent Variable Model.

KEYWORDS IN $L_2$: .

*Proceedings of COLING 2012: Technical Papers*, pages 2805–2820,
COLING 2012, Mumbai, December 2012.

## 1 Introduction

Recognizing Textual Entailment (RTE) (Dagan et al., 2005) is the task of recognizing entailment relations between a given text pair, *Text T* and *Hypothesis H*. RTE is useful for many information access tasks that depend on natural language processing technologies, and a breakthrough would lead to significant progress in information retrieval, document summarization, and question answering, among other tasks.

The majority of approaches proposed in previous work recognize entailment relations between a pair of texts by capturing lexical or structural correspondences. Methods include simple word overlap-based measures (Jijkoun and de Rijke, 2005) as well as alignment of syntactic and semantic dependencies (Sammons et al., 2009; Wang and Zhang, 2009). However, sentence-level semantic relations are affected by various linguistic phenomena: not only lexical semantic relations (synonyms, antonyms) but also monotonicity (e.g. downward-monotone caused by scope of negation), implicative/factive expressions, quantifiers, etc. Thus similarity measures are insufficient to capture these phenomena and their interactions.

Transformation-based approaches are one way to capture the affects of diverse linguistic phenomena and their interactions, where a set of linguistic phenomena are decomposed into units. By doing so it becomes possible to consider their effects on entailment independently. A number of previous works explores transformation-based entailment relation recognition. The approach of Stern et al. (2011) recognizes a sentence-level semantic relation through a proof which represents a sequence of edits from $T$ to $H$ produced by applying various entailment rules and the operations such as insertion, deletion, moving subtrees, etc. In addition, Heilman and Smith (2010) proposed a tree edit model which selects a sequence of edits using Tree Kernels, and Wang and Manning (2010) proposed a latent variable model which consider possible alignments as hidden structures. However, these model do not sufficiently represent interactions between linguistic phenomena such as factuality reversals caused by negation and flipping of entailment direction under downward-monotone contexts. In order to realize precise entailment relation recognition, we need to appropriately deal with semantic relations resulting from the interaction between linguistic phenomena.

One of the most promising approaches to RTE is Natural Logic-based recognition (MacCartney and Manning, 2008; MacCartney, 2009). This approach represents transformations from $T$ to $H$ with a set of three types of alignment edits (*substitution*, *insertion* and *deletion*), and assigns one of a set-theoretically defined semantic relations to each alignment edit. This approach is based on the principle of compositionality, i.e. the sentence-level semantic relation is derived by combining semantic relations of edits using pre-defined composition rules. By doing so, this approach makes progress toward precise sentence-level entailment relation recognition that considers linguistic phenomena and their interactions when assigning semantic relations.

However, several issues remain unexplored. While it is common for alignment inference methods to require data annotated with alignments, it is a challenge to manually annotate alignments in a consistent manner. Annotation of alignments with semantic relations from Natural Logic is a greater challenge due to the complex nature of the semantic relations. In addition, even alignments can be annotated consistently, there is no guarantee of their *global optimality*; that is to say the alignments identified as correct by annotators may not necessarily contribute to identifying the correct semantic relation between a pair of sentences. Identifying alignments considering the full context of a sentence pair is a much more difficult annotation task. However, even without manual alignment annotations, it may be possible to infer consistent and plausible alignments by learning models that promote alignments which agree with annotations of correct semantic relations between sentences. A unified model of alignment and semantic relation

recognition between sentences is needed that learns the alignments which will generate the correct semantic relation by considering the interaction between diverse linguistic phenomena.

In this paper, we propose a novel latent discriminative model that jointly handles predicting alignment edits, classification of their semantic relations and entailment relation recognition by providing a joint distribution of variables including alignment edits, their local semantic relations and sentence-level semantic relations. Inspired by the Natural Logic-based approach of (MacCartney et al., 2008), we incorporate the set of semantic relations and their composition rules from Natural Logic into our proposed model. In addition, our model can be trained from only sentence-level semantic relations to predict alignments and semantic relations that are consistent with Natural Logic composition. To the best of our knowledge, our study is the first work to propose a latent model for training a Natural Logic-based semantic relation recognition system that does not require alignment annotations and that jointly predicts plausible alignments and semantic relations between sentences, modeling a variety of linguistic phenomena and their interactions in a compositional manner.

## 2 Natural Logic

The concept of *Natural Logic*, a logic over natural language, is originally proposed by Lakoff (1970), and then van Benthem (1988, 1991) and Valencia (1991) explored monotonicity calculus[1] to explain entailment relations using Natural Logic. While they considered only containment relations, MacCartney and Manning (2008) introduced an *exclusion* relation to deal with entailment relations which involve different objects or concepts (e.g. Stimpy is a cat $\models$ Stimpy is not a poodle). In this section, we describe the theory of Natural Logic proposed by (MacCartney and Manning, 2008; MacCartney, 2009).

The basic idea of MacCartney et al's theory is that the semantic relation between sentences can be derived from the semantic relations of *edits* (substitution, deletion and insertion) from $T$ to $H$. The fundamental assumption of the theory is *compositionality*: (some of) the entailments of a compound expression are a function of the entailments of its parts. They defined the seven types of semantic relations for edits: equivalence ($a \equiv b$ if $a = b$), forward-entailment ($a \sqsubset b$ if $a \subset b$), backward-entailment ($a \sqsupset b$ if $b \supset a$), negation ($a \,^\wedge\, b$ if $a \cap b = \phi \wedge a \cup b = U$)[2], alternation ($a \mid b$ if $a \cap b = \phi \wedge a \cup b \neq U$), cover ($a \cup b \neq \phi \wedge a \cup b = U$), and independence ($a \# b$ otherwise).

Semantic relations provided by edits are *projected* onto other relations depending on their contexts using *projection rules*. For example, in a scope of negation, forward-entailment is projected onto backward-entailment (e.g *soccer* $\sqsubset$ *sports*, *I didn't play soccer.* $\sqsupset$ *I didn't play sports.*). Other linguistic expressions such as logical connectives and quantifiers also projects semantic relations. A semantic relation between sentences is derived by combining the projected semantic relations of edits using *composition rules*. The rules are defined as tuples of semantic relations. Let the seven types of relations be $\mathscr{R}$, $r_i \in \mathscr{R}$, $r_j \in \mathscr{R}$, then a compositional rule is represented by $r_i \bowtie r_j \Rightarrow \boldsymbol{r} \subseteq \mathscr{R}$. Some compositional rule derive a single relations (e.g. $\equiv \bowtie \sqsubset \Rightarrow \sqsubset$), and others derive more than one semantic relations (e.g. $\mid \bowtie \mid \Rightarrow \bigcup \{\equiv, \sqsubset, \sqsupset, \mid, \#\}$). As semantic relation composition proceeded, semantic relations tend to move toward $\#$ [3].

[1] In an *upward-monotone* context, replacing a linguistic expression with a more general expression preserves truth. On the other hand, in a *downward-monotone* context, replacing a linguistic expression with a more specific expression preserves truth.

[2] $U$ denotes a universe.

[3] Due to spacial limitations, we can not give all of the composition rules. For more details, see (MacCartney, 2009).

## 3 A Latent Discriminative Model for Compositional Entailment Relation Recognition

Given a text $T$ and a hypothesis $H$, the task of RTE is to infer the correct semantic relation between $T$ and $H$. However, we attempt to learn not only the correct semantic relation between $T$ and $H$ but also the characteristics of the alignments most likely to support that relation.

We assume that sentence-level semantic relations can be derived compositionally. Following the framework of Natural Logic proposed by (MacCartney and Manning, 2008; MacCartney, 2009), our proposed model assigns local semantic relations to edits which represent a transformation from $T$ to $H$. A valid set of edits represents an alignment between $T$ and $H$. Each edit is categorized as one of three types: *substitution*, *deletion* or *insertion*, and is given one of the seven semantic relations defined in Natural Logic described in §2. A semantic relation between $T$ and $H$ is derived from a set of semantic relations of alignment edits by using the projection rules and the composition rules.

The proposed model learns appropriate alignments which are consistent with compositional rules of Natural Logic from only sentence-level semantic relations, where appropriate alignments, their semantic relations and their projections are represented using hidden variables. We use a log-linear discriminative model with hidden variables to provide conditional joint probabilities of alignments, their associated semantic relations, and their projections and a sentence-level semantic relation.

### 3.1 Model

Our proposed model provides a conditional joint distribution of alignment edits, their semantic relations, their projected relations and the final semantic relation between $T$ and $H$ as follows.

$$p(\boldsymbol{e}, \boldsymbol{r_e}, \boldsymbol{r_e}^P, \boldsymbol{r}^C | \boldsymbol{x}; \boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{x})} \exp\left( \sum_k \Psi_k(\boldsymbol{e}, \boldsymbol{r_e}, \boldsymbol{r_e}^P, \boldsymbol{r}^C, \boldsymbol{x}; \boldsymbol{\lambda}) \right) \quad (1)$$

$\boldsymbol{e} = \{e_i\}$ denotes the variables representing edits, and each edit $e_i = \langle \boldsymbol{t}_i, \boldsymbol{h}_i \rangle$ consists of $\boldsymbol{t}_i$, a subset of indices of units (e.g. words) in $T$, and $\boldsymbol{h}_i$, a subset of indices of units in $H$. An edit corresponds to substitution if $\boldsymbol{t}_i \neq \phi$ and $\boldsymbol{h}_i \neq \phi$, deletion if $\boldsymbol{t}_i \neq \phi$ and $\boldsymbol{h}_i = \phi$, and insertion if $\boldsymbol{t}_i = \phi$ and $\boldsymbol{h}_i \neq \phi$. $\boldsymbol{r_e}$ represents the set of semantic relations for $\boldsymbol{e}$, where $r_{e_i} \in \boldsymbol{r_e}$ corresponds to the semantic relation of $e_i$. Since $r_{e_i}$ is derived without considering its context, $r_{e_i}$ can be seen as the semantic relation between $\boldsymbol{t}_i$ and $\boldsymbol{h}_i$. The variables $\boldsymbol{r_e}^P$ represents a set of projected semantic relations derived from $r_{\boldsymbol{e}}$, taking into account their contexts. If an edit is under the scope of negation, a quantifier or a conditional, then $r_{e_i}$ is mapped to an appropriate semantic relation $r_{e_i}^P$ based on that context. Therefore $r_{e_i}^P$ can be seen as the sentence-level semantic relation between $T$ and the sentence which can be obtained by applying the edit $e_i$ to $T$. The variables $\boldsymbol{r}^C$ denotes a set of semantic relations derived by combining $\boldsymbol{r_e}^P$, where each $r^C \in \boldsymbol{r}^C$ corresponds to the result of composition of two semantic relations. Hereafter, we use $r_T^C$ as the sentence-level semantic relation. Note that $r_T^C \in \boldsymbol{r}^C$. Each variable r in $\boldsymbol{r_e}$, $\boldsymbol{r_e}^P$ and $\boldsymbol{r}^C$ can have seven types of semantic relations described previously. $\Psi_k$ in equation (1) is a *factor* which scores the plausibility of alignment edits, their semantic relations, etc.

Our proposed model uses the following four types of factors to score the plausible alignment edits, their semantic relations and a sentence-level semantic relation.

**Alignment Factor** $\Psi_A(e, \boldsymbol{x})$ is used to deal with (unlabeled) phrase alignment for entailment relation recognition and is defined as $\Psi_A(e, \boldsymbol{x}) = \boldsymbol{\lambda} \cdot \boldsymbol{f}_A(e, \boldsymbol{x})$. In order to provide good alignments, it is necessary to capture the lexical similarity between words. The features used in this factor are mainly (i) surface-based similarity between alignment units, (ii) semantic relatedness of alignment units, which can be extracted from diverse lexical knowledge databases, and (iii) the contextual information for an edit.

**Alignment Semantic Relation Factor** $\Psi_S(e, r_e, \boldsymbol{x})$ is introduced to provide plausibility of a semantic relation $r_e \in \boldsymbol{r_e}$ for an alignment edit $e \in \boldsymbol{e}$ and is defined as $\Psi_S = \boldsymbol{\lambda} \cdot \boldsymbol{f}_S(e, r_e, \boldsymbol{x})$. Each variable $r_e$ has a distribution over the seven types of semantic relations defined in Natural Logic. In order to classify semantic relations, not only surface-based similarities, but also lexical semantic relations play an important role. In the NatLog system developed by (MacCartney, 2009), an implementation of an RTE system of Natural Logic, lexical resource-derived features (e.g. WordNet, NomBank, etc.), string similarity features, and lexical category features are used. For this factor, we exploit diverse lexical resources to provide informative features for classifying semantic relations of edits.

**Projection Factor** $\Psi_P(r_e, r_e^P, \boldsymbol{x})$ provides an appropriate projection from $r_e$ to $r_e^P$ by considering the context of $e$, and is defined by $\Psi_P(r_e, r_e^P, \boldsymbol{x}) = \boldsymbol{\lambda} \cdot \boldsymbol{f}_P(r_e, r_e^P, \boldsymbol{x})$. This factor captures the effects of monotonicity (e.g. upward, downward). Given $r_e$ and its contexts, the semantic relation of the projected variable $r_e^P$ is uniquely determined using the monotonicity rules of (MacCartney, 2009).

**Composition Factor** $\Psi_C(r_{i-1}^C, r_e^P, r_i^C, \boldsymbol{x})$ scores tuples of semantic relations, and is defined by $\Psi_{C_i}(r_{i-1}^C, r_e^P, r_i^C, \boldsymbol{x}) = \boldsymbol{\lambda} \cdot \boldsymbol{f}(r_{i-1}^C, r_e^P, r_i^C, \boldsymbol{x})$. In this factor, we use the composition rules used in (MacCartney, 2009) with some modification. We set the derived semantic relations to independence (#) for the rules which derive more than one semantic relations. Therefore, as with $\Psi_P$, given two semantic relations $r_{i-1}^C$ and $r_e^P$, the joined relation of the variable $r_i^C$ is uniquely determined.

An overview of the proposed model is shown in Figure 1. In this figure, we show the factor graph constructed by our proposed model for a pair of sentences in Japanese. Our model is divided by three layers: the *alignment layer*, the *projection layer* and the *composition layer*. First, in the alignment layer, our proposed model scores possible alignments using $\Psi_A$ and $\Psi_S$. For alignment units, we use *bunsetsu* which is a reasonable unit for Japanese linguistic analysis. A *bunsetsu* is a chunk-like unit that consists of one or more content words and zero or more functional words. A set of possible alignments are obtained using an extended MANLI algorithm (MacCartney et al., 2008).

Next, for each alignment obtained by the alignment algorithm, we construct a factor graph as shown in Figure 1. The factor graph has variables for alignments, projected relations, joined relations, and the factors defined previously. In the projection layer, semantic relations of alignments are projected by $\Psi_P$, and finally a sentence-level semantic relation is obtained in the composition layer using the projected relations and composition rules encoded in $\Psi_C$.

In inference, since variables related to $\Psi_P$ and $\Psi_C$ are uniquely determined if $\boldsymbol{r_e}$ is given, the model derives the best alignments, their semantic relations, and a sentence-level semantic relation simultaneously. In training, the parameters of the model are updated so as to derive alignments and their semantic relations which derive the correct sentence-level semantic relation based on the composition rules of Natural Logic.

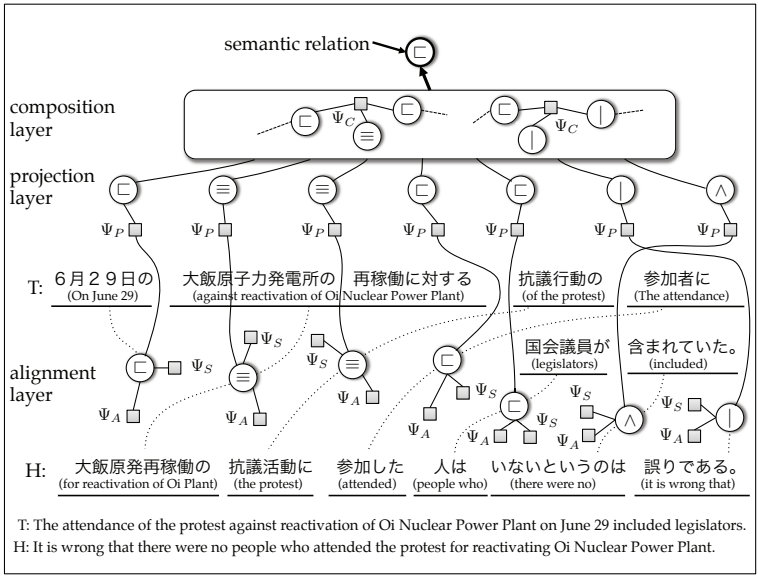


Figure 1: An overview of the proposed model.

## 3.2 Features

The features used in the proposed model are listed in Table 3.2.

Because RTE datasets are small, it is difficult to incorporate lexical features directly into our model as they may cause overfitting. Instead, we incorporate similarity metrics to model lexicality. On the other hand, because function words are closed class and present in all texts, we can directly use them as features.

While both $\Psi_A$ and $\Psi_S$ score the plausibility of alignments, the features used in their factors are also different. $\Psi_A$ considers not only lexical similarities but also contexts of edits. Let us consider a simple sentence pair *T: USA won the war but Japan lost the war. H: Japan won the war.* In this example, T and H share the same verb *won* but the word *won* in H should be aligned to *lose* in T because they share the same subject (*Japan*). So, we introduce features that capture predicate-argument structure-level contextual information: e.g. how many arguments are shared by the two predicates (NUM_SHARED_ARGS)? On the other hand, $\Psi_S$ pays more attention to inferring the lexical semantic relations of edits. The features used in $\Psi_p$ and $\Psi_C$ work as rules to infer sentence-level semantic relations.

## 3.3 Learning the Model

The parameters $\boldsymbol{\lambda}$ of the proposed model are trained from sentence-level semantic relations via marginal-likelihood maximization $\mathscr{L}_{\boldsymbol{\lambda}} = \sum_n \log p(r_T^C = l^n | \boldsymbol{x}^n; \boldsymbol{\lambda})$. By applying this objective function, we expect that the proposed model is trained so as to prefer alignment edits and

| Factor | Edit | Name | Description |
|---|---|---|---|
| $\Psi_A$ | DEL, INS | TYPE | edit type of $e$ |
| | | SIZE | the number of *bunsetsus* in $e$ |
| | | SAME_NOMINATIVE_IN_{T,H} | 1 if $e$ has a nominative argument and the other sentence also contains a nominative argument and its lemmas are the same. |
| | | SAME_CASE_IN_{T,H} | 1 if $e$ has an argument of some predicates and the other sentence also contains an argument and its cases are the same. |
| | | {T,H}_CONTAINS_{H,T}_LEMMA | 1 if the head word of the *bunsetsu* in $e$ is also contained in the other sentence. |
| | | POS_SEQ | POS sequence in $e$ |
| | | HEADPOS | head POS of $e$ |
| $\Psi_A$ | SUB | TYPE | edit type of $e$ |
| | | SIZE | the number of *bunsetsus* in $e$ |
| | | NUM_SHARED_ARGS | the number of shared arguments if both $t$ and $h$ in $e$ are predicates |
| | | PARTICLE_SAME | 1 if $t$ and $t$ have the same particle |
| | | JAPANESE_WORDNET | Set relation type if $e$ matches an entry in Japanese WordNet (Bond et al., 2009). |
| | | WIKIPEDIA_HYPERNYM-HYPONYM | 1 if $e$ matches an entry in (Sumida et al., 2008). |
| | | VERB_ENTAILMENT_REL | 1 if an entry in the verb entailment relation dictionary (Hashimoto et al., 2009) matches $e$ |
| | | VERB_RELATION_REL | 1 if an entry in the verb relation dictionary (Matsuyoshi et al., 2008) matches $e$ |
| | | PARENT_NUM_SHARED_ARGS | the number of shared arguments of the parent of $t$ and $h$ if $t$ and $h$ are arguments |
| | | BOTH_HAVE_A_ROLE | 1 if each *bunsetsu* in $e$ is an argument of a predicate. |
| | | BOTH_HAVE_THE_SAME_ROLE | 1 if each *bunsetsu* in $e$ has the same case. |
| | | HEAD_POS_SAME | 1 if the POSs of the heads in $e$ are the same |
| | | POS_SEQ_SAME | 1 if the POS sequences of chunks in $e$ are the same |
| | | EXACT_MATCH | 1 if $t$ and $h$ are the same |
| | | UNIGRAM_COSINE | return unigram cosine value if the cosine similarity of two chunks in $e$ is greater than the pre-defined threshold |
| $\Psi_S$ | DEL, INS | SIZE | number of *bunsetsus* in $e$ |
| | | HEAD_LEMMA | lemma of the head of *bunsetsu* in $e$ |
| | | HEAD_WORD_CLASS | Word class of the head of *bunsetsu* in $e$. The word class information is extracted from the dictionary provided by (Kazama et al., 2010) |
| | | NEGATION | 1 if the *bunsetsu* contains a negation |
| $\Psi_S$ | SUB | SIZE | Pair of the number of *bunsetsus* in $e$ |
| | | HEAD_POS_PAIR | POS pair of the heads of *bunsetsus* in $e$ |
| | | HEAD_LEMMA_SAME | 1 if the lemmas of the heads in $e$ are the same |
| | | POS_SEQ_SAME | 1 if the POS sequences of chunks in $e$ are the same |
| | | JAPANESE_WORDNET | same as in $\Psi_A$ |
| | | WIKIPEDIA_HYPERNYM-HYPONYM | same as in $\Psi_A$ |
| | | VERB_ENTAILMENT_REL | same as in $\Psi_A$ |
| | | VERB_RELATION_REL | same as in $\Psi_A$ |
| $\Psi_P$ | – | MONOTONE_{UP/DOWN} | the context of $e$ is *upward-monotone* or *downward-monotone*. |
| $\Psi_C$ | – | COMPOSITION_RULE | 1 if the tuple of semantic relations is included in a set of defined compositional rules. |

Table 1: Features used for the model.

their semantic relations which infer the correct sentence-level semantic relations based on the composition rules of Natural Logic.

The partial differential of the objective function is

$$\frac{\partial L}{\partial \lambda_k} = \sum_n \left( \frac{\partial}{\partial \lambda_k} \log \sum_{\langle \boldsymbol{e}, \boldsymbol{r}_e \rangle \in \mathcal{E}} \sum_{\boldsymbol{r}: r_T^C = l} \exp\left( \sum_k \boldsymbol{\Psi}_k(\boldsymbol{e}, \boldsymbol{r}_e, \boldsymbol{r}_e^P, \boldsymbol{r}^C, \boldsymbol{x}) \right) - \frac{\partial}{\partial \lambda_k} \log Z(\boldsymbol{x}) \right) \quad (2)$$

**Algorithm 1** The alignment algorithm.

**input**
an example $(\boldsymbol{x}_T, \boldsymbol{x}_H)$, number of iterations $I$, max size of edits $M$,
number of N-bests $N$, score difference $\delta$, score function $\Psi_A(\boldsymbol{e}, \boldsymbol{x}) + \Psi_S(\boldsymbol{e}, \boldsymbol{r}_e, \boldsymbol{x})$
**initialize**
$\boldsymbol{e}_0 \leftarrow \phi$
$\forall x_T \in \boldsymbol{x}_T\ \boldsymbol{e}_0 \leftarrow \boldsymbol{e}_0 \cup e(x_T, DEL, \sqsubset)$
$\forall x_H \in \boldsymbol{x}_H\ \boldsymbol{e}_0 \leftarrow \boldsymbol{e}_0 \cup e(x_H, INS, \sqsupset)$
all alignments $\mathscr{E} \leftarrow \boldsymbol{e}_0$
**while** $(iter < I)$ **do**
  top alignments $\mathscr{E}_{top}$
  max score $score_{argmax}$
  **repeat**
    get top alignment $\boldsymbol{e}_{top}$ from $\mathscr{E}$
    **if** $s(\boldsymbol{e}_{top}) > score_{argmax}$ **then**
      $score_{argmax} \leftarrow s(\boldsymbol{e}_{top})$
    **end if**
    $\mathscr{E}_{top} \leftarrow \mathscr{E}_{top} \cup \boldsymbol{e}_{top}$
  **until** $score_{argmax} - s(\boldsymbol{e}_{top}) >= \delta$
  get successors $\mathscr{S}_i = \{\boldsymbol{e}_i^s\}_i$ for $\boldsymbol{e}_i \in \mathscr{E}_{top}$
  $\mathscr{E} \leftarrow \mathscr{E} \bigcup_i \mathscr{S}_i$
**end while**
**return** $\mathscr{E}$

where edits $\boldsymbol{e}$, their semantic relations $\boldsymbol{r}_e$, projected semantic relations $\boldsymbol{r}_e^P$ and joined relations excluding the sentence-level semantic relation are all hidden variables. Given $\boldsymbol{r}_e$, $\boldsymbol{r}_e^P$ and $\boldsymbol{r}^C$ can be identified uniquely by using projection and composition rules. Since the objective function is non-convex, estimated parameters can be local-optima.

In optimization, only the parameters in $\Psi_A$ and $\Psi_S$ are updated, and the parameters in $\Psi_P$ and $\Psi_C$ are left to initial values. In order to update the parameters, we need to calculate marginal probabilities of the alignments. However, unlike sequential or tree models, calculating exact values of alignments is prohibitively difficult. We use only N-bests provided by the extended MANLI algorithm to calculate an approximate partition function $\widetilde{Z}(\boldsymbol{x})$ instead of $Z(\boldsymbol{x})$, and approximate marginal probabilities.

## 3.4 Inference of Alignments

Given two sentences, the problem of alignment inference in our model is predicting the best edits and their semantic relations $\widehat{\langle \boldsymbol{e}, \boldsymbol{r}_e \rangle} = \arg\max_{\langle \boldsymbol{e}, \boldsymbol{r}_e \rangle \in \mathscr{E}} \sum_{\langle e_i, r_{e_i} \rangle \in \langle \boldsymbol{e}, \boldsymbol{r}_e \rangle} \Psi_A(e_i, \boldsymbol{x}; \boldsymbol{\lambda}) + \sum_{\langle e_i, r_{e_i} \rangle \in \langle \boldsymbol{e}, \boldsymbol{r}_e \rangle} \Psi_S(e_i, r_{e_i}, \boldsymbol{x}; \boldsymbol{\lambda})$ where $\mathscr{E}$ is a set of all possible edits and their semantic relations between two sentences. The original MANLI algorithm (MacCartney et al., 2008) only provides the best edits, so we extend the algorithm so as to provide not only edits, but also their semantic relations.

The extended version of MANLI is shown in Algorithm 1. Given two sentences, the algorithm starts at an initial alignment $\boldsymbol{e}_0$ which consists of deletion edits of *bunsetsus* in $T$ and insertion edits of *bunsetsu* in $H$, and then searches for more good alignments by changing edits from a pair of a deletion and an insertion edit to a substitution edit, or changing semantic labels. The main differences between the original MANLI and our algorithm are: (1) alignments have their semantic relations, (2) keeps a set of alignments ordered by scores provided by $\Psi_A$ and $\Psi_S$ to provide N-bests. We omitted the annealing procedure which is included in the original

MANLI because our algorithm need to keep an ordered set of alignments based on scores. If we introduce a temperature value, we have to update all of the alignments in the set when the value is changed. However this is computationally expensive.

### 3.5 The Order of Composition

The composition order of semantic relations defined in Natural Logic is non-commutative. Let us consider joining an alternation ($|$) and a forward-entailment ($\sqsubset$). The pair of semantic relations frequently appear in contradiction examples. $\sqsubset$ joined with $|$ yields $|$, on the other hand, $|$ joined with $\sqsubset$ yields $\bigcup\{\equiv, ^\wedge, |, \cup, \#\}$. The former way of composition derives the desired result, however, the latter way derives an ambiguous result. We defined the order of composition so as to keep joined semantic relations unambiguous as far as possible. Our proposed model at first joins $\equiv$ and $\sqsubset$, then $|$, then $^\wedge$, and $\sqsupset$ in the end [4].

## 4 Experiments

### 4.1 Data

We developed a dataset for semantic relation recognition which includes a diverse selection of linguistic phenomena. Although there is an textual entailment recognition data set for Japanese (RITE (Shima et al., 2011)), we do not consider it an appropriate target for evaluation and instead construct our own dataset. Our motivation is as follows. Much of the progress made in textual entailment recognition has been on a set of phenomena that can be handled with methods of lexical and phrasal similarity, however, there are many other phenomena that have not been addressed.

Sammons et al. (2010) make a case for more detailed analysis of the linguistic phenomena important to textual entailment recognition so that their impact on existing approaches can be properly measured. In that spirit, we investigated textual entailment recognition phenomena and found that quantification, negation, and monotonicity require consideration of their semantic structure and are beyond the scope of similarity-based methods. Constructing systematic and robust models of handling these phenomena is the focus of this paper. It is reasonable to target these phenomena next because many of the remaining problems for textual entailment recognition require world knowledge and are thus problems of inference or AI. Existing datasets for textual entailment recognition are insufficient for our purposes because the phenomena they contain are too broad and they do not contain enough examples of the phenomena we are targeting to draw meaningful conclusions.

We selected the categories based on FraCaS (Cooper et al., 1996), the corpus developed by Bentivogli et al. (2010) and the categories discussed in (MacCartney, 2009): lexical semantic relation (e.g. synonym, antonym, hypernym-hyponym relation), quantifiers, modifiers, negation, coordination, relative clauses, apposition, temporal and numerical expressions, active/passive, factive verbs and functional relations.

The statistics of the dataset is shown in Table 4.1. The distribution of the categories is not balanced: the quantifier category accounts for approximately 30% of the total. One of our interests is whether the model can automatically capture behaviour of functional expressions such as quantifiers from sentence-level semantic relations. In order to conform this point, we developed many examples for quantifiers.

[4] NatLog uses a different strategy from ours. The system at first joins semantic relations of deletion edits, then substitution edits, next edits involve operators with non-default projectivity, and, finally, insertion edits.

| Category | # | Category | # | Category | # |
|---|---|---|---|---|---|
| Quantifier | 182 | Coordination | 27 | Part-Whole | 13 |
| Numerical/Temporal | 53 | Argument Mismatch | 26 | Condition | 8 |
| Modifier | 55 | Negation + Lexical semantic Rel. | 23 | Apposition | 7 |
| Lexical semantic relation | 44 | Paraphrase | 21 | World Knowledge | 3 |
| Implicative/Factive | 36 | Predicate Mismatch | 17 | Other | 37 |
| Relation between entities/events | 27 | Coordination | 27 | Total | 598 |

Table 2: Category statistics.

| Category | Example | Sem. Rel. |
|---|---|---|
| Quantifier | *T: Almost all mammals have molars in the back of their rows of teeth.*<br>*H: There are mammals that do not have molars.* | Forward-Ent. |
| Paraphrase | *T: Not all smokers get cancer.*<br>*H: Even if you smoke, you might not get cancer.* | Paraphrase |
| Modifier | *T: There aren't many students.*<br>*H: There aren't many students who have had a heat stroke.* | Forward-Ent. |
| Lexical semantic rel. | *T: Japan got a bronze medal in Team Fencing.*<br>*H: Japan hasn't gotten a bronze medal in any sports.* | Contradiction |
| Implicative/Factive | *T: Earthquake-proofing prevented the house's collapse.*<br>*H: The house did not collapse.* | Forward-Ent. |
| Coordination | *T: Tokyo has a population of 13,000,000 and Miyagi has a population of 2,300,000.*<br>*H: Tokyo has a population of 2,300,000.* | Contradiction |

Table 3: Some examples in the dataset (translated in English).

For each example, we annotated one of the four types of sentence-level semantic relations (paraphrase, forward-entailment, contradiction and independence), and alignment edits and their semantic relations in Natural Logic. In the dataset, the number of paraphrase examples is 97, forward-entailment is 313, contradiction is 100, and independence is 88. Table 4.1 shows some examples in the dataset. The dataset was developed by one annotator, who is a professionally trained linguist unaffiliated with this research project, and the set of annotated semantic relations does not always provide the correct semantic relation. 55.2% of the gold annotations derive correct sentence-level semantic relation (332 examples). The remaining examples include inconsistencies between sentence-level semantic relations and semantic relations of alignments, linguistic phenomena that the current model can not deal with (e.g. syntactic transformation, some quantifiers), etc.

Whereas there are seven types of relations in Natural Logic, our annotation uses only four types of relations. So in the experiments, we mapped *contradiction* to $\{\wedge, |\}$ and *other* to $\{\cup, \#\}$ in the training and the testing phase.

## 4.2 Settings

In order to explore the effectiveness of the proposed model, we evaluated the following approaches in the experiments.

**Initial Weight** Initial weights of the model are used for testing.

**Resource-based Alignment** Alignments are determined based on a surface-based similarity measures and lexical resources. In this setting, a pair of two phrases is aligned if the character-bigram cosine similarity is greater than a pre-defined threshold (we set it to 0.8), or the pair matches an entry in the lexical resources such as Japanese WordNet, Hypernym-Hyponym relations, Verb Entailment Relations, and Verb Relation Dictionary.

Semantic relations of alignment edits are determined as follows: $\equiv$ if the pair of the *bunsetsus* $\langle b_1, b_2 \rangle$ is the same, similar or is synonym, | if the pair is antonym, $\sqsubset$ if $b_2$ is the hypernym of $b_1$ or $b_1$ entails $b_2$, $\sqsupset$ if $b_1$ is the hypernym of $b_2$ or $b_2$ entails $b_1$. The *bunsetsus* not aligned by the similarity measure or the resources are converted to *deletion* or *insertion* edits, and their semantic relations are set to $\sqsubset$ and $\sqsupset$ respectively with exceptions described later.

**Alignment Supervised** The model is trained using gold alignments which have correct semantic relations defined in Natural Logic. In this setting, sentence-level semantic relations are not considered in training. As in the proposed model, we constructed the model using a log-linear discriminative model, and the model was trained log-likelihood maximization of gold alignments. The objective function used in training was $\mathcal{L}_{\boldsymbol{\lambda}} = \sum_n \log p(\boldsymbol{e}, \boldsymbol{r}_e | \boldsymbol{x}^n; \boldsymbol{\lambda})$.

**Weakly Supervised (proposed)** The model is trained by marginal likelihood maximization over sentence-level semantic relations.

The dataset we used in the experiments include the examples whose correct sentence-level semantic relations can not be derived from the pre-annotated semantic relations of alignment edits. It seems that these are hard to derive correct sentence-level semantic relations from the current possible edits. So, we conducted experiments on the examples whose correct sentence-level semantic relations can be derived from the gold alignments (hereafter, we say *reachable*).

For the factors $\Psi_P$ and $\Psi_C$, we initialized the weights to 0.0 if the semantic relation tuple is covered by our projection rules and composition rules, and $-\infty$ otherwise. For the factors $\Psi_A$ and $\Psi_S$, we set initial weights to some features [5]. In training of the model, we update the parameters in $\Psi_A$ and $\Psi_S$, and the parameters in $\Psi_P$ and $\Psi_C$ are left to the initial values. Parameter updating was performed using stochastic gradient descent (SGD), and the number of iterations was set to 2. Also, we applied $L_2$ regularization. As for the alignment algorithm, the number of iterations was set to 40, and the number of N-bests was set to 10. For each edit type, we restricted the maximum size of units: only allows one-to-one for substitution, allows at most three units for insertion and deletion edits. Also, we constrained the types of semantic relations for each edit type. Substitution edits can have one of the five types of semantic relations: $\equiv$, $\sqsubset$, $\sqsupset$, $\wedge$ and | with an exception. If the lemma sequences of the two *bunsetsus* are the same, the edit can have only $\equiv$. Deletion edits and insertion edits can have $\sqsubset$ and $\sqsupset$ respectively with exceptions. They can have | if the head of *bunsetsu* matches an entry in the list of *counter-factive expressions* [6], and they can have $\equiv$ if the head of *bunsetsu* matches an entry in the list of *less-informative expressions* [7].

## 4.3 Evaluation Measures

We use the following measures in evaluation: **(1) Alignment (Unlabeled):** A predicted alignment is correct if there is a gold alignment which has the same span, but the semantic label is not considered, **(2) Alignment (Labeled):** A predicted alignment is correct only if there is a gold alignment which has the same span and their semantic relations are also the same, and **(3) Sem. Rel.:** Accuracy of sentence-level semantic relations.

[5] For instance, the weights of the combination feature "NEGATION=0" and "JAPANESE_WORDNET=antonym" are set to 1.0 if label is | and $-1.0$ otherwise

[6] A hand-crafted list which contains 13 entries.

[7] As with the list of *counter-factive expressions*, the list was hand-crafted, and contains 30 entries.

| | Alignment (Unlabeled) | | | Alignment (Labeled) | | | Sem. Rel. |
|---|---|---|---|---|---|---|---|
| | Prec. | Rec. | F1 | Prec. | Rec. | F1 | Acc. |
| Initial Weights | 42.6 | 62.5 | 50.6 | 37.5 | 54.9 | 44.5 | 31.8 |
| Resource-based Alignment | 45.6 | 63.3 | 53.0 | 38.6 | 53.5 | 44.9 | 35.0 |
| Weakly Supervised (proposed) | 67.1 | 67.8 | 67.4 | 51.3 | 51.9 | 51.6 | **47.5** |
| Alignment Supervised | 68.0 | 68.8 | **68.4** | 54.9 | 55.5 | **55.2** | 43.7 |
| Gold Alignment | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 55.2 |
| reachable examples only | | | | | | | |
| Initial Weights | 45.6 | 65.5 | 53.8 | 41.7 | 59.9 | 49.2 | 38.5 |
| Resource-based Alignment | 48.4 | 66.7 | 56.1 | 42.6 | 58.8 | 49.4 | 37.7 |
| Weakly Supervised (proposed) | 72.4 | 73.1 | 72.7 | 59.2 | 59.8 | 59.5 | **60.2** |
| Alignment Supervised | 74.2 | 75.0 | **74.6** | 61.9 | 62.6 | **62.3** | 46.4 |

Table 4: Performance of alignment prediction and sentence-level semantic relation recognition.

## 4.4 Preprocessing

For each sentence, we conducted various forms of linguistic analysis: morphological analysis using MeCab (Kudo et al., 2004), syntactic parsing using the Japanese dependency parser, CaboCha (Kudo and Matsumoto, 2002) and predicate-argument structure analysis (Watanabe et al., 2010) to provide a basis for alignment and semantic relation classification.

## 4.5 Results

Table 4 shows the experimental results of 10-fold cross validation for alignment prediction and sentence level semantic relation recognition. We can see that while the proposed method is less successful at reproducing gold standard alignments, it greatly outperforms Supervised Learning for sentence-level semantic relation recognition[8]. We expected Supervised Learning to perform best on reachable examples, which should have the most straightforward connection between alignment semantic relation labels and sentence level semantic relations. Nevertheless, our proposed method achieved the best performance on this dataset as well. These results support our theory that gold standard alignment data is necessary for semantic relation recognition. Indeed, alignment labels appear to degrade performance in several cases.

Table 5 shows the sentence-level performance for each semantic relation type. This breakdown shows that the proposed method is particularly good at Contradiction and Forward-Entailment relations, outperforming all other methods on all data sets. When considering reachable examples only, it is also the top-performing method for Paraphrase detection as well. Resource-based Alignment and Initial Weights both perform poorly, producing significantly worse results than the supervised methods in every evaluation setting with the exception of Contradiction on reachable examples only and Independence on both data sets.

The poor performance by Resource-based Alignment and Initial Weights is likely due to inaccurate alignments, especially of functional expressions (e.g. *sometimes - not always*). Since deletion and an insertion edits are assigned $\sqsubset$ and $\sqsupset$ respectively by default and joining them yields independence (#), these methods over-produce Independence relations. Most of the errors in Alignment Supervised are caused by lower precision for alternation (|). Since alternation relations can greatly impact the sentence-level semantic relation prediction, this severely impacted the overall performance of the supervised model.

Table 6 shows the performances of semantic relation classification of alignments for each type. As discussed before, supervised alignment is the most successful at recovering gold standard

[8] We compared the sentence-level semantic relation recognition results of Weakly Supervised and Alignment Supervised with the McNemar test, and the difference was statistically significant ($p < 0.01$).

| | Paraphrase | | | Forward-Entailment | | |
|---|---|---|---|---|---|---|
| | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| Resource-based Alignment | 67.4 | 29.9 | 41.4 | 73.6 | 28.4 | 41.0 |
| Initial Weights | 61.1 | 11.3 | 19.1 | 81.9 | 27.5 | 41.2 |
| Weakly Supervised (proposed) | 46.1 | 54.6 | 50.0 | 73.9 | 55.3 | **63.2** |
| Alignment Supervised | 60.5 | 47.4 | **53.2** | 72.7 | 39.9 | 51.6 |
| | Contradiction | | | Independence | | |
| Resource-based Alignment | 19.2 | 15.0 | 16.9 | 22.1 | 86.4 | 35.2 |
| Initial Weights | 41.4 | 12.0 | 18.6 | 18.6 | 92.1 | 31.0 |
| Weakly Supervised (proposed) | 25.0 | 43.0 | **31.6** | 33.3 | 17.1 | 22.6 |
| Alignment Supervised | 21.3 | 42.0 | 28.3 | 36.4 | 54.6 | **43.6** |
| reachable examples only | | | | | | |
| | Paraphrase | | | Forward-Entailment | | |
| Resource-based Alignment | 52.9 | 20.9 | 30.0 | 76.3 | 33.7 | 46.8 |
| Initial Weights | 60.0 | 20.9 | 31.0 | 85.7 | 34.9 | 49.6 |
| Weakly Supervised (proposed) | 59.5 | 51.2 | **55.0** | 62.3 | 85.5 | **72.1** |
| Alignment Supervised | 53.1 | 39.5 | 45.3 | 61.2 | 58.7 | 59.9 |
| | Contradiction | | | Independence | | |
| Resource-based Alignment | 29.8 | 20.6 | 24.4 | 23.2 | 89.8 | 36.8 |
| Initial Weights | 63.2 | 17.7 | 27.6 | 20.6 | 95.9 | 33.9 |
| Weakly Supervised (proposed) | 73.9 | 25.0 | **37.4** | 60.9 | 28.6 | 38.9 |
| Alignment Supervised | 19.0 | 26.5 | 22.1 | 62.1 | 36.7 | **46.2** |

Table 5: The details of the results of sentence-level entailment relation recognition.

| | Equivalence (≡) | | | Forward-Entailment (⊏) | | |
|---|---|---|---|---|---|---|
| | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| Resource Alignment | 60.7 | 76.0 | 67.5 | 25.5 | 49.8 | 33.8 |
| Initial Weights | 69.0 | 76.2 | 72.4 | 22.8 | 53.3 | 31.9 |
| Weakly Supervised | 66.0 | 89.0 | 75.8 | 30.1 | 25.3 | 27.5 |
| Alignment Supervised | 70.0 | 89.5 | 78.6 | 35.2 | 29.3 | 32.0 |
| | Backward-Entailment (⊐) | | | Alternation (\|), Negation (^) | | |
| Resource Alignment | 11.8 | 47.0 | 18.1 | 68.4 | 14.3 | 23.7 |
| Initial Weights | 9.7 | 54.8 | 16.4 | 91.2 | 5.6 | 10.5 |
| Weakly Supervised | 27.1 | 13.9 | 18.3 | 28.7 | 15.4 | 20.0 |
| Alignment Supervised | 34.3 | 15.1 | 20.9 | 23.3 | 17.2 | 19.8 |
| reachable examples only | | | | | | |
| | Equivalence (≡) | | | Forward-Entailment (⊏) | | |
| | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| Resource Alignment | 69.1 | 76.5 | 72.6 | 27.4 | 53.8 | 36.3 |
| Initial Weights | 75.3 | 76.9 | 76.1 | 25.2 | 57.0 | 35.0 |
| Weakly Supervised | 72.8 | 88.7 | 80.0 | 30.5 | 37.5 | 33.6 |
| Alignment Supervised | 76.8 | 89.5 | 82.7 | 35.1 | 38.0 | 36.5 |
| | Backward-Entailment (⊐) | | | Alternation (\|), Negation (^) | | |
| | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| Resource Alignment | 8.3 | 52.5 | 14.3 | 70.6 | 19.1 | 30.0 |
| Initial Weights | 7.2 | 57.6 | 12.8 | 100.0 | 8.3 | 15.4 |
| Weakly Supervised | 25.0 | 11.9 | 16.1 | 83.3 | 7.9 | 14.5 |
| Alignment Supervised | 28.1 | 15.3 | 19.8 | 40.3 | 19.8 | 26.6 |

Table 6: The details of the performances of alignment prediction.

alignments and semantic relation labels. However, it is interesting to note that while Resource Alignment performs competitively at alignment prediction (it rivals Alignment Supervised on Forward-Entailment and outperforms all other methods on Alternation/Negation), it performs drastically worse on sentence-level semantic relation recognition, sometimes with an f-score that is more than 20 points lower than the best performing method. These results suggest that it is important to jointly model alignment prediction and sentence-level semantic relation recognition so that globally optimal alignments are promoted.

## 5 Related Work

There are a number of existing works which explore the use of latent variable or structure models for recognizing textual entailment. Chang et al. (2010) proposed a discriminative linear model where alignments are treated as hidden structures, and the sentence-level semantic relation is derived based on the best latent alignment structure. They formulated the problem of predicting the best hidden structure as an Integer Linear Programming problem, where domain knowledge is encoded as constraints. Wang and Manning (2010) proposed a latent variable model where the model provides a conditional distribution of a sequence of edits, which can be seen as a transformation-based approach. In the model, edits are treated as hidden variables that populate a positive set and a negative set in the search space. Sentence-level semantic relations are predicted based on the sum of the scores of edit sequences in the positive set and the negative set.

The differences between our proposed model and theirs are that the number of semantic relations and compositionality. Both Wang and Manning (2010) and Chang et al. (2010) consider only *entailment* and *non-entailment*, while our proposed model identifies a rich set of relations: *paraphrase*, *forward entailment*, *backward entailment*, *contradiction*, and *independence*.

Furthermore, as discussed in Section 2, our model exhibits compositionality by incorporating Natural Logic at two different levels. First, it incorporates information about upward and downward monotonicity into a projection layer, allow it to handle flips in entailment direction caused by scope of negation that can influence the final sentence-level semantic relation. In addition, it considers the result of combining projected semantic relations of alignment edits, allowing it to handle complex interactions between linguistic phenomena in sentences. The alignment models of Wang and Manning (2010) and Chang et al. (2010) do not consider the interaction between alignments that we model with Natural Logic making it difficult for them to classify examples that contain complex semantic structures.

## Conclusion

In this paper, we proposed a novel latent variable model for compositional entailment relation recognition. We gave the proposed model compositionality by incorporating a set of semantic relations and their composition rules of Natural Logic. The model has ability to predict local correspondences (alignments) between sentences, the semantic relations, and the sentence-level semantic relation simultaneously. The model can be trained from only sentence-level semantic relations by using marginal-likelihood maximization. In evaluation, our proposed method outperformed a supervised alignment method on a sentence-level semantic relation recognition task, and detailed analysis on that task and an alignment prediction task did not provide evidence that gold standard alignment labels contributed to semantic relation recognition performance and instead suggests that they can be detrimental to performance if used in a manner that prevents the learning of globally optimal alignments.

A future research direction we are investigating is extending the model so as to deal with structural transformations. The current model has a big drawback: the model assumes that all of sentence-level semantic relations can be derived from only *bunsetsu* -level transformations. We would like to explore how to incorporate transformation rules (used in e.g. (Stern et al., 2011)) into the proposed model.

## Acknowledgements

This work is supported by the PRESTO program of JST and the Grants-in-Aid for Scientific Research No. 23240018, No. 23700157 and No. 23700159.

## References


Bentivogli, L., Cabrio, E., Dagan, I., Giampiccolo, D., Leggio, M. L., and Magnini, B. (2010). Building textual entailment specialized data sets: a methodology for isolating linguistic phenomena relevant to inference. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., and Kanzaki, K. (2009). Enhancing the Japanese WordNet. In *Proceedings of the 7th Workshop on Asian Language Resources, ACL-IJCNLP 2009*, pages 1–8.

Chang, M., Goldwasser, D., Roth, D., and Srikumar, V. (2010). Discriminative learning over constrained latent representations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 429–437.

Cooper, R., Crouch, D., van Eijck, J., Fox, C., van Genabith, J., Jaspers, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., and Pulman, S. (1996). Using the framework. Technical report, Technical Report LRE 62-051 D-16, The FraCaS Consortium.

Dagan, I., Glickman, O., and Magnini, B. (2005). The Pascal Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*.

Hashimoto, C., Torisawa, K., Kuroda, K., Murata, M., and Kazama, J. (2009). Large-scale verb entailment acquisition from the web. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP 2009)*, pages 1172–1181.

Heilman, M. and Smith, N. (2010). Tree Edit Models for Recognizing Textual Entailments, Paraphrases, and Answers to Questions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1011–1019.

Jijkoun, V. and de Rijke, M. (2005). Recognizing textual entailment using lexical similarity. In *Proceedings of the PASCAL Challenges Workshop on RTE*.

Kazama, J., Saeger, S. D., Kuroda, K., Murata, M., and Torisawa, K. (2010). A bayesian method for robust estimation of distributional similarities. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Kudo, T. and Matsumoto, Y. (2002). Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pages 63–69.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Appliying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Lakoff, G. (1970). George Lakoff - Linguistics and Natural Logic. *Synthese*, 22:151–271.

MacCartney, B. (2009). *Natural Language Inference*. PhD thesis.

MacCartney, B., Galley, M., and Manning, C. (2008). A Phrase-Based Alignment Model for Natural Language Inference. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 802–811.

MacCartney, B. and Manning, C. D. (2008). Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 521–528.

Matsuyoshi, S., Murakami, K., Matsumoto, Y., , and Inui, K. (2008). A database of relations between predicate argument structures for recognizing textual entailment and contradiction. In *Proceedings of the 2nd International Symposium on Universal Communication (ISUC2008)*, pages 366–373.

Sammons, M., Vydiswaran, V., and Roth, D. (2010). Ask not what textual entailment can do for you... In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1199–1208.

Sammons, M., Vydiswaran, V., Vieira, T., Johri, N., Chang, M., Goldwasser, D., Srikumar, V., Kundu, G., Tu, Y., and Small, K. (2009). Relation alignment for textual entailment recognition.

Shima, H., Kanayama, H., Lee, C.-W., Lin, C.-J., Mitamura, T., Miyao, Y., Shi, S., and Takeda, K. (2011). Overview of NTCIR-9 RITE: Recognizing Inference in TExt. In *Proceedings of NTCIR-9 Workshop Meeting*, pages 291–301.

Stern, A., Lotan, A., Mirkin, S., Shnarch, E., Kotlerman, L., Berant, J., and Dagan, I. (2011). Knowledge and Tree-Edits in Learnable Entailment Proofs. In *Proceedings of the Fourth Text Analysis Conference (TAC 2011)*.

Sumida, A., Yoshinaga, N., and Torisawa, K. (2008). Boosting Precision and Recall of Hyponymy Relation Acquisition from Hierarchical Layouts in Wikipedia. In *Proceedings of the 6th International Language Resources and Evaluation (LREC'08)*, pages 2462–2469.

Valencia, V. S. (1991). *Studies on Natural Logic and Categorial Grammar*. PhD thesis.

van Benthem, J. (1988). The semantics of variety in categorial grammars. pages 33–55.

van Benthem, J. (1991). Language in action: Categories, lambdas and dynamic logic.

Wang, M. and Manning, C. (2010). Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. pages 1164–1172.

Wang, R. and Zhang, Y. (2009). Recognizing textual relatedness with predicate-argument structures. In *EMNLP-2009*.

Watanabe, Y., Asahara, M., and Matsumoto, Y. (2010). A structured model for joint learning of argument roles and predicate senses. In *Proceedings of the 48th Annual Meeting of the Association of Computational Linguistics*, pages 98–102.