LEDIR: An Unsupervised Algorithm for Learning Directionality of Inference Rules

Rahul Bhagat, Patrick Pantel, Eduard Hovy
Information Sciences Institute
University of Southern California
Marina del Rey, CA
{rahul,pantel,hovy}@isi.edu

Abstract

Semantic inference is a core component of many natural language applications. In response, several researchers have developed algorithms for automatically learning inference rules from textual corpora. However, these rules are often either imprecise or underspecified in directionality. In this paper we propose an algorithm called LEDIR that filters incorrect inference rules and identifies the directionality of correct ones. Based on an extension to Harris's distributional hypothesis, we use selectional preferences to gather evidence of inference directionality and plausibility. Experiments show empirical evidence that our approach can classify inference rules significantly better than several baselines.

1 Introduction

Paraphrases are textual expressions that convey the same meaning using different surface forms. Textual entailment is a similar phenomenon, in which the presence of one expression licenses the validity of another. Paraphrases and inference rules are known to improve performance in various NLP applications like Question Answering (Harabagiu and Hickl 2006), summarization (Barzilay et al. 1999) and Information Retrieval (Anick and Tipirneni 1999).

Paraphrase and entailment involve inference rules that license a conclusion when a premise is given. Deciding whether a proposed inference rule is fully valid is difficult, however, and most NL systems instead focus on plausible inference. In this case, one statement has some likelihood of being identical in meaning to, or derivable from, the other. In the rest of this paper we discuss plausible inference only.

Given the importance of inference, several researchers have developed inference rule collections. While manually built resources like WordNet (Fellbaum 1998) and Cyc (Lenat 1995) have been around for years, for coverage and domain adaptability reasons many recent approaches have focused on automatic acquisition of paraphrases (Barzilay and McKeown 2001) and inference rules (Lin and Pantel 2001; Szpektor et al. 2004). The downside of these approaches is that they often result in incorrect inference rules or in inference rules that are underspecified in directionality (i.e. asymmetric but are wrongly considered symmetric). For example, consider an inference rule from DIRT (Lin and Pantel 2001):

$$X\ eats\ Y \Leftrightarrow X\ likes\ Y \qquad (1)$$

All rules in DIRT are considered symmetric. Though here, one is most likely to infer that "*X eats Y*" ⇒ "*X likes Y*", because if someone eats something, he most probably likes it[1], but if he likes something he might not necessarily be able to eat it. So for example, given the sentence "*I eat spicy food*", one is mostly likely to infer that "*I like spicy food*". On the other hand, given the sentence "*I like rollerblading*", one cannot infer that "*I eat rollerblading*".

In this paper, we propose an algorithm called **LEDIR** (pronounced "leader") for **LE**arning **D**irectionality of **I**nference **R**ules. Our algorithm filters incorrect inference rules and identifies the directionality of the correct ones. Our algorithm

[1] There could be certain usages of "*X eats Y*" where, one might not be able to infer "*X likes Y*" (for example metaphorical). But, in most cases, this inference holds.

works with any resource that produces inference rules of the form shown in example (1). We use both the distributional hypothesis and selectional preferences as the basis for our algorithm. We provide empirical evidence to validate the following main contribution:

***Claim**: Relational selectional preferences can be used to automatically determine the plausibility and directionality of an inference rule.*

2 Related Work

In this section, we describe applications that can benefit by using inference rules and their directionality. We then talk about some previous work in this area.

2.1 Applications

Open domain question answering approaches often cast QA as the problem of finding some kind of semantic inference between a question and its answer(s) (Moldovan et al. 2003; Echiabi and Marcu 2003). Harabagiu and Hickl (2006) recently demonstrated that textual entailment inference information, which in this system is a set of directional inference relations, improves the performance of a QA system significantly even without using any other form of semantic inference. This evidence supports the idea that learning the directionality of other sets of inference rules may improve QA performance.

In Multi-Document Summarization (MDS), paraphrasing is useful for determining sentences that have similar meanings (Barzilay et al. 1999). Knowing the directionality between the inference rules here could allow the MDS system to choose either the more specific or general sentence depending on the purpose of the summary.

In IR, paraphrases have been used for query expansion, which is known to promote effective retrieval (Anick and Tipirneni 1999). Knowing the directionality of rules here could help in making a query more general or specific depending on the user needs.

2.2 Learning Inference Rules

Automatically learning paraphrases and inference rules from text is a topic that has received much attention lately. Barzilay and McKeown (2001) for paraphrases, DIRT (Lin and Pantel 2001) and TEASE (Szpektor et al. 2004) for inference rules, are recent approaches that have achieved promising results. While all these approaches produce collections of inference rules that have good recall, they suffer from the complementary problem of low precision. They also make no attempt to distinguish between symmetric and asymmetric inference rules. Given the potential positive impact shown in Section 2.1 of learning the directionality of inference rules, there is a need for methods, such as the one we present, to improve existing automatically created resources.

2.3 Learning Directionality

There have been a few approaches at learning the directionality of restricted sets of semantic relations, mostly between verbs. Chklovski and Pantel (2004) used lexico-syntactic patterns over the Web to detect certain types of symmetric and asymmetric relations between verbs. They manually examined and obtained lexico-syntactic patterns that help identify the types of relations they considered and used these lexico-syntactic patterns over the Web to detect these relations among a set of candidate verb pairs. Their approach however is limited only to verbs and to specific types of verb-verb relations.

Zanzotto et al. (2006) explored a selectional preference-based approach to learn asymmetric inference rules between verbs. They used the selectional preferences of a single verb, i.e. the semantic types of a verb's arguments, to infer an asymmetric inference between the verb and the verb form of its argument type. Their approach however applies also only to verbs and is limited to some specific types of verb-argument pairs.

Torisawa (2006) presented a method to acquire inference rules with temporal constraints, between verbs. They used co-occurrences between verbs in Japanese coordinated sentences and co-occurrences between verbs and nouns to learn the verb-verb inference rules. Like the previous two methods, their approach too deals only with verbs and is limited to learning inference rules that are temporal in nature.

Geffet and Dagan (2005) proposed an extension to the distributional hypothesis to discover entailment relation between words. They model the context of a word using its syntactic features and compare the contexts of two words for strict inclusion to infer lexical entailment. In principle, their work is the most similar to ours. Their method however is limited to lexical entailment and they show its effectiveness for nouns. Our method on the other hand deals with inference rules between binary relations and includes inference rules between verbal relations, non-verbal relations and multi-word relations. Our definition of context and the methodology for obtaining context similarity and overlap is also much different from theirs.

3 Learning Directionality of Inference Rules

The aim of this paper is to filter out incorrect inference rules and to identify the directionality of the correct ones.

Let $p_i \Leftrightarrow p_j$ be an inference rule where each *p* is a binary semantic relation between two entities *x* and *y*. Let <*x*, *p*, *y*> be an instance of relation *p*.

Formal problem definition: *Given the inference rule* $p_i \Leftrightarrow p_j$*, we want to conclude which one of the following is more appropriate:*

1. $p_i \Leftrightarrow p_j$
2. $p_i \Rightarrow p_j$
3. $p_i \Leftarrow p_j$
4. *No plausible inference*

Consider the example (1) from section 1. There, it is most plausible to conclude "*X eats Y*" ⇒ "*X likes Y*".

Our algorithm LEDIR uses selectional preferences along the lines of Resnik (1996) and Pantel et al. (2007) to determine the plausibility and directionality of inference rules.

3.1 Underlying Assumption

Many approaches to modeling lexical semantics have relied on the distributional hypothesis (Harris 1954), which states that words that appear in the same contexts tend to have similar meanings. The idea is that context is a good indicator of a word meaning. Lin and Pantel (2001) proposed an extension to the distributional hypothesis and applied it to paths in dependency trees, where if two paths tend to occur in similar contexts it is hypothesized that the meanings of the paths tend to be similar.

In this paper, we assume and propose a further extension to the distributional hypothesis and call it the "Directionality Hypothesis".

Directionality Hypothesis: *If two binary semantic relations tend to occur in similar contexts and the first one occurs in significantly more contexts than the second, then the second most likely implies the first and not vice versa.*

The intuition here is that of generality. The more general a relation, more the types (and number) of contexts in which it is likely to appear. Consider the example (1) from section 1. The fact is that there are many more things that someone might like than those that someone might eat. Hence, by applying the directionality hypothesis, one can infer that "*X eats Y*" ⇒ "*X likes Y*".

The key to applying the distributional hypothesis to the problem at hand is to model the contexts appropriately and to introduce a measure for calculating context similarity. Concepts in semantic space, due to their abstractive power, are much richer for reasoning about inferences than simple surface words. Hence, we model the context of a relation *p* of the form <*x*, *p*, *y*> by using the semantic classes *C(x)* and *C(y)* of words that can be instantiated for *x* and *y* respectively. To measure context similarity of two relations, we calculate the overlap coefficient (Manning and Schütze, 1999) between their contexts.

3.2 Selectional Preferences

The selectional preferences of a predicate is the set of semantic classes that its arguments can belong to (Wilks 1975). Resnik (1996) gave an information theoretical formulation of the idea. Pantel et al. (2007) extended this idea to non-verbal relations by defining the relational selectional preferences (RSPs) of a binary relation *p* as the set of semantic classes *C*(*x*) and *C*(*y*) of words that can occur in positions *x* and *y* respectively.

The set of semantic classes *C*(*x*) and *C*(*y*) can be obtained either from a manually created taxonomy like WordNet as proposed in the above previous approaches or by using automatically generated classes from the output of a word clustering algorithm as proposed in Pantel et al. (2007). For example given a relation like "*X likes Y*", its RSPs from WordNet could be *{individual, social_group…}* for *X* and *{individual, food, activity…}* for *Y*.

In this paper, we deployed both the Joint Relational Model (JRM) and Independent Relational Model (IRM) proposed by Pantel et al. (2007) to obtain the selectional preferences for a relation *p*.

Model 1: Joint Relational Model (JRM)

The JRM uses a large corpus to learn the selectional preferences of a binary semantic relation by considering its arguments jointly.

Given a relation *p* and large corpus of English text, we first find all occurrences of relation *p* in the corpus. For every instance <*x, p, y*> in the corpus, we obtain the sets *C*(*x*) and *C*(*y*) of the semantic classes that *x* and *y* belong to. We then accumulate the frequencies of the triples <*c*(*x*), *p*, *c*(*y*)> by assuming that every $c(x) \in C(x)$ can co-occur with every $c(y) \in C(y)$ and vice versa. Every triple <*c*(*x*), *p*, *c*(*y*)> obtained in this manner is a candidate selectional preference for *p*. Following Pantel et al. (2007), we rank these candidates using Pointwise mutual information (Cover and Thomas 1991). The ranking function is defined as the strength of association between two semantic classes, c_x and c_y[2], given the relation *p*:

$$pmi\left(c_x|p; c_y|p\right) = \log \frac{P\left(c_x, c_y|p\right)}{P\left(c_x|p\right)P\left(c_y|p\right)} \tag{3.1}$$

Let $|c_x, p, c_y|$ denote the frequency of observing the instance <*c*(*x*), *p*, *c*(*y*)>. We estimate the probabilities of Equation 3.1 using maximum likelihood estimates over our corpus:

$$P\left(c_x|p\right) = \frac{|c_x, p, *|}{|*, p, *|} \quad P\left(c_y|p\right) = \frac{|*, p, c_y|}{|*, p, *|} \qquad P\left(c_x, c_y|p\right) = \frac{|c_x, p, c_y|}{|*, p, *|} \tag{3.2}$$

We estimate the above frequencies using:

$$|c_x, p, *| = \sum_{w \in c_x} \frac{|w, p, *|}{|C(w)|} \quad |*, p, c_y| = \sum_{w \in c_y} \frac{|*, p, w|}{|C(w)|} \qquad |c_x, p, c_y| = \sum_{w_1 \in c_x, w_2 \in c_y} \frac{|w_1, p, w_2|}{|C(w_1)| \times |C(w_2)|} \tag{3.3}$$

where |*x*, *p*, *y*| denotes the frequency of observing the instance <*x*, *p*, *y*> and |C(*w*)| denotes the number of classes to which word *w* belongs. |C(*w*)| distributes *w*'s mass equally among all of its senses C(*w*).

Model 2: Independent Relational Model (IRM)

Due to sparse data, the JRM is likely to miss some pair(s) of valid relational selectional preferences. Hence we use the IRM, which models the arguments of a binary semantic relation independently.

Similar to JRM, we find all instances of the form <*x, p, y*> for a relation *p*. We then find the sets *C*(*x*) and *C*(*y*) of the semantic classes that *x* and *y* belong to and accumulate the frequencies of the triples <*c*(*x*), *p*, *> and <*, *p*, *c*(*y*)> where $c(x) \in C(x)$ and $c(y) \in C(y)$.

All the tuples <*c*(*x*), *p*, *> and <*, *p*, *c*(*y*)> are the independent candidate RSPs for a relation *p* and we rank them according to equation 3.3.

Once we have the independently learnt RSPs, we need to convert them into a joint representation for use by the inference plausibility and directionality model. To do this, we obtain the Cartesian product between the sets <*C*(*x*), *p*, *> and <*, *p*, *C*(*y*)> for a relation *p*. The Cartesian product between two sets A and B is given by:

$$A \times B = \left\{(a,b) : \forall a \in A \quad and \quad \forall b \in B\right\} \tag{3.4}$$

Similarly we obtain:

$$\langle C_x, p, * \rangle \times \langle *, p, C_y \rangle = \left\{ \begin{matrix} \langle c_x, p, c_y \rangle : \quad \forall \langle c_x, p, * \rangle \in \langle C_x, p, * \rangle \quad and \\ \forall \langle *, p, c_y \rangle \in \langle *, p, C_y \rangle \end{matrix} \right\} \tag{3.5}$$

The Cartesian product in equation 3.5 gives the joint representation of the RSPs of the relation *p* learnt using IRM. In the joint representation, the IRM RSPs have the form <*c*(*x*), *p*, *c*(*y*)> which is the same form as the JRM RSPs.

3.3 Inference plausibility and directionality model

Our model for determining inference plausibility and directionality is based on the intuition that for an inference to hold between two semantic relations there must exist sufficient overlap between their contexts and the directionality of the inference depends on the quantitative comparison between their contexts.

Here we model the context of a relation by the selectional preferences of that relation. We determine the plausibility of an inference based on the overlap coefficient (Manning and Schütze, 1999) between the selectional preferences of the two paths. We determine the directionality based on the difference in the number of selectional preferences of the relations when the inference seems plausible.

Given a candidate inference rule $p_i \Leftrightarrow p_j$, we first obtain the RSPs <*C*(*x*), p_i, *C*(*y*)> for p_i and <*C*(*x*), p_j, *C*(*y*)> for p_j. We then calculate the overlap coefficient between their respective RSPs. Overlap coefficient is one of the many distribu-

[2] c_x and c_y are shorthand for *c*(*x*) and *c*(*y*) in our equations.

tional similarity measures used to calculate the similarity between two vectors A and B:

$$sim(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)} \qquad (3.6)$$

The overlap coefficient between the selectional preferences of p_i and p_j is calculated as:

$$sim(p_i,p_j) = \frac{|\langle C_x,p_i,C_y\rangle \cap \langle C_x,p_j,C_y\rangle|}{\min(|C_x,p_i,C_y|,|C_x,p_j,C_y|)} \qquad (3.7)$$

If sim(p_i,p_j) is above a certain empirically determined threshold α (≤1), we conclude that the inference is plausible, i.e.:

```
If  $sim(p_i,p_j) \geq \alpha$
        we conclude the inference is plausible
else
        we conclude the inference is not plausible
```

For a plausible inference, we then compute the ratio between the number of selectional preferences $|C(x), p_i, C(y)|$ for p_i and $|C(x), p_j, C(y)|$ for p_j and compare it against an empirically determined threshold β (≥1) to determine the direction of inference. So the algorithm is:

```
If       $\frac{|C_x,p_i,C_y|}{|C_x,p_j,C_y|} \geq \beta$        we conclude $p_i \Leftarrow p_j$
else if  $\frac{|C_x,p_i,C_y|}{|C_x,p_j,C_y|} \leq \frac{1}{\beta}$     we conclude $p_i \Rightarrow p_j$
else                                  we conclude $p_i \Leftrightarrow p_j$
```

4 Experimental Setup

In this section, we describe our experimental setup to validate our claim that LEDIR can be used to determine plausibility and directionality of an inference rule.

Given an inference rule of the form $p_i \Leftrightarrow p_j$, we want to use automatically learned relational selectional preferences to determine whether the inference rule is valid and if it is valid then what its directionality is.

4.1 Inference Rules

LEDIR can work with any set of binary semantic inference rules. For the purpose of this paper, we chose the inference rules from the DIRT resource (Lin and Pantel 2001). DIRT consists of 12 million rules extracted from 1GB of newspaper text (AP Newswire, San Jose Mercury and Wall Street Journal). For example, "*X eats Y*" ⇔ "*X likes Y*" is an inference rule from DIRT.

4.2 Semantic Classes

Appropriate choice of semantic classes is crucial for learning relational selectional preferences. The ideal set should have semantic classes that have the right balance between abstraction and discrimination, the two important characteristics that are often conflicting. A very general class has limited discriminative power, while a very specific class has limited abstractive power. Finding the right balance here is a separate research problem of its own.

Since the ideal set of universally acceptable semantic classes in unavailable, we decided to use the Pantel et al. (2007) approach of using two sets of semantic classes. This approach gave us the advantage of being able to experiment with sets of classes that vary a lot in the way they are generated but try to maintain the granularity by obtaining approximately the same number of classes.

The first set of semantic classes was obtained by running the CBC clustering algorithm (Pantel and Lin, 2002) on TREC-9 and TREC-2002 newswire collections consisting of over 600 million words. This resulted in 1628 clusters, each representing a semantic class.

The second set of semantic classes was obtained by using WordNet 2.1 (Fellbaum 1998). We obtained a cut in the WordNet noun hierarchy[3] by manual inspection and used all the synsets below a cut point as the semantic class at that node. Our inspection showed that the synsets at depth four formed the most natural semantic classes[4]. A cut at depth four resulted in a set of 1287 semantic classes, a set that is much coarser grained than WordNet which has an average depth of 12. This seems to be a depth that gives a reasonable abstraction while maintaining good discriminative power. It would however be interesting to experiment with more sophisticated algorithms for extracting semantic classes from WordNet and see their effect

[3] Since we are dealing with only noun binary relations, we use only WordNet noun Hierarchy.

[4] By natural, here, we simply mean that a manual inspection by the authors showed that, at depth four, the resulting clusters had struck a better granularity balance than other cutoff points. We acknowledge that this is a very coarse way of extracting concepts from WordNet.

on the relational selectional preferences, something we do not address this in this paper.

4.3 Implementation

We implemented LEDIR with both the JRM and IRM models using inference rules from DIRT and semantic classes from both CBC and WordNet. We parsed the 1999 AP newswire collection consisting of 31 million words with Minipar (Lin 1993) and used this to obtain the probability statistics for the models (as described in section 3.2).

We performed both system-wide evaluations and intrinsic evaluations with different values of α and β parameters. Section 5 presents these results and our error analysis.

4.4 Gold Standard Construction

In order to evaluate the performance of the different systems, we compare their outputs against a manually annotated gold standard. To create this gold standard, we randomly sampled 160 inference rules of the form $p_i \Leftrightarrow p_j$ from DIRT. We discarded three rules since they contained nominalizations[5].

For every inference rule of the form $p_i \Leftrightarrow p_j$, the annotation guideline asked annotators (in this paper we used two annotators) to choose the most appropriate of the four options:

1. $p_i \Leftrightarrow p_j$
2. $p_i \Rightarrow p_j$
3. $p_i \Leftarrow p_j$
4. *No plausible inference*

To help the annotators with their decisions, the annotators were provided with 10 randomly chosen instances for each inference rule. These instances, extracted from DIRT, provided the annotators with context where the inference could hold. So for example, for the inference rule "*X eats Y*" ⇔ "*X likes Y*", an example instance would be "*I eat spicy food"* ⇔ "*I like spicy food*". The annotation guideline however gave the annotators the freedom to think of examples other than the ones provided to make their decisions.

The annotators found that while some decisions were quite easy to make, the more complex ones often involved the choice between bi-directionality and one of the directions. To minimize disagreements and to get a better understanding of the task, the annotators trained themselves by annotating several samples together.

We divided the set of 157 inference rules, into a development set of 57 inference rules and a blind test set of 100 inference rules. Our two annotators annotated the development test set together to train themselves. The blind test set was then annotated individually to test whether the task is well defined. We used the kappa statistic (Siegel and Castellan Jr. 1988) to calculate the inter-annotator agreement, resulting in κ=0.63. The annotators then looked at the disagreements together to build the final gold standard.

All this resulted in a final gold standard of 100 annotated DIRT rules.

4.5 Baselines

To get an objective assessment of the quality of the results obtained by using our models, we compared the output of our systems against three baselines:

B-random: *Randomly assigns one of the four possible tags to each candidate inference rule.*

B-frequent: *Assigns the most frequently occurring tag in the gold standard to each candidate inference rule.*

B-DIRT: *Assumes each inference rule is bidirectional and assigns the bidirectional tag to each candidate inference rule.*

5 Experimental Results

In this section, we provide empirical evidence to validate our claim that the plausibility and directionality of an inference rule can be determined using LEDIR.

5.1 Evaluation Criterion

We want to measure the effectiveness of LEDIR for the task of determining the validity and directionality of a set of inference rules. We follow the standard approach of reporting system accuracy by comparing system outputs on a test set with a manually created gold standard. Using the gold standard described in Section 4.4, we measure the accuracy of our systems using the following formula:

[5] For the purpose of simplicity, we in our experiments did not use DIRT rules containing nominalizations. The algorithm however can be applied without change to inference rules containing nominalization. In fact, in the resource that we plan to release soon, we have applied the algorithm without change to DIRT rules containing nominalizations.

$$Accuracy = \frac{|correctly \quad tagged \quad \inf erences|}{|input \quad \inf erences|}$$

5.2 Result Summary

We ran all our algorithms with different parameter combinations on the development set (the 57 DIRT rules described in Section 4.4). This resulted in a total of 420 experiments on the development set. Based on these experiments, we used the accuracy statistic to obtain the best parameter combination for each of our four systems. We then used these parameter values to obtain the corresponding percentage accuracies on the test set for each of the four systems.

Model		α	β	Accuracy (%)
B-random		-	-	25
B-frequent		-	-	34
B-DIRT		-	-	25
JRM	CBC	0.15	2	38
	WN	0.55	2	38
IRM	**CBC**	**0.15**	**3**	**48**
	WN	0.45	2	43

Table 1: Summary of results on the test set

Table 1 summarizes the results obtained on the test set for the three baselines and for each of the four systems using the best parameter combinations obtained as described above. The overall best performing system uses the IRM algorithm with RSPs form CBC. Its performance is found to be significantly better than all the three baselines using the Student's paired t-test (Manning and Schütze, 1999) at $p<0.05$. However, this system is not statistically significant when compared with the other LEDIR implementations (JRM and IRM with WordNet).

5.3 Performance and Error Analysis

The best performing system selected using the development set is the IRM system using CBC with the parameters $\alpha=0.15$ and $\beta=3$. In general, the results obtained on the test set show that the IRM tends to perform better than the JRM. This observation points at the sparseness of data available for learning RSPs for the more restrictive JRM, the reason why we introduced the IRM in the first place. A much larger corpus would be needed to obtain good enough coverage for the JRM.

		GOLD STANDARD			
		⇔	⇒	⇐	NO
SYSTEM	⇔	16	1	3	7
	⇒	0	3	1	3
	⇐	7	4	22	15
	NO	2	3	4	9

Table 2: Confusion Matrix for the best performing system, IRM using CBC with $\alpha=0.15$ and $\beta=3$.

Table 2 shows the confusion matrix for the overall best performing system as selected using the development set (results are taken from the test set). The confusion matrix indicates that the system does a very good job of identifying the directionality of the correct inference rules, but gets a big performance hit from its inability to identify the incorrect inference rules accurately. We will analyze this observation in more detail below.

Figure 1 plots the variation in accuracy of IRM with different RSPs and different values of α and β. The figure shows a very interesting trend. It is clear that for all values of β, systems for IRM using CBC tend to reach their peak in the range $0.15 \leq \alpha \leq 0.25$, whereas the systems for IRM using WordNet (WN), tend to reach their peak in the range $0.4 \leq \alpha \leq 0.6$. This variation indicates the kind of impact the selection of semantic classes could have on the overall performance of the system. This is not hard evidence, but it does suggest that finding the right set of semantic classes could be one big step towards improving system accuracy.

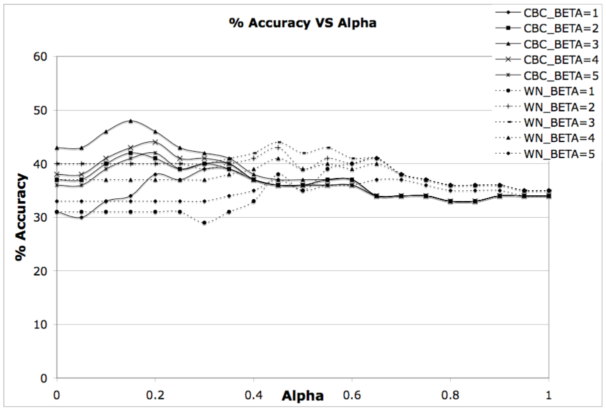

Figure 1: Accuracy variation for IRM with different values of α and β.

Two other factors that have a big impact on the performance of our systems are the values of the system parameters α and β, which decide the plausibility and directionality of an inference rule, respectively. To better study their effect on the system performances, we studied the two parameters independently.

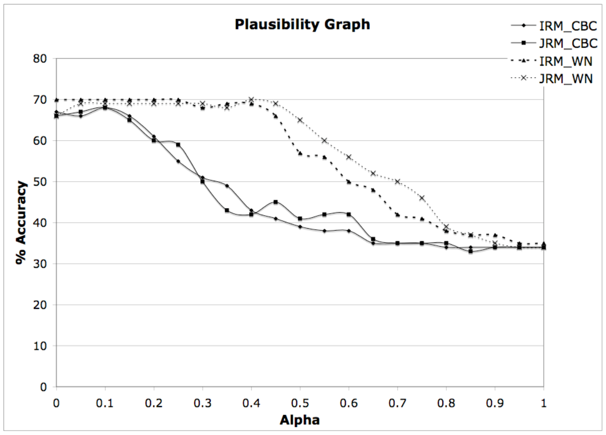

Figure 2: Accuracy variation in predicting correct versus incorrect inference rules for different values of α.

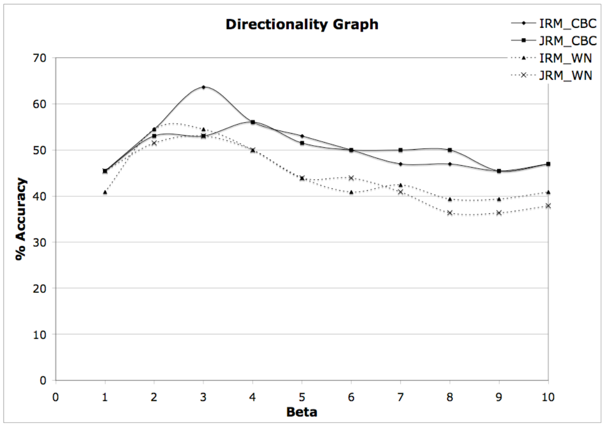

Figure 3: Accuracy variation in predicting directionality of correct inference rules for different values of β.

Figure 2 shows the variation in the accuracy for the task of predicting the correct and incorrect inference rules for the different systems when varying the value of α. To obtain this graph, we classified the inference rules in the test set only as correct and incorrect without further classification based on directionality. All of our four systems obtained accuracy scores in the range of 68-70% showing a good performance on the task of determining plausibility. This however is only a small improvement over the baseline score of 66% obtained by assuming every inference to be plausible (as will be shown below, our system has most impact not on determining plausibility but on determining directionality). Manual inspection of some system errors showed that the most common errors were due to the well-known 'problem of antonymy' when applying the distributional hypothesis. In DIRT, one can learn rules like "*X loves Y*" ⇔ "*X hates Y*". Since the plausibility of inference rules is determined by applying the distributional hypothesis and the antonym paths tend to take the same set of classes for X and Y, our models find it difficult to filter out the incorrect inference rules which DIRT ends up learning for this very same reason. To improve our system, one avenue of research is to focus specifically on filtering incorrect inference rules involving antonyms (perhaps using methods similar to (Lin et al. 2003)).

Figure 3 shows the variation in the accuracy for the task of predicting the directionality of the correct inference rules for the different systems when varying the value of β. To obtain this graph, we separated the correct inference rules form the incorrect ones and ran all the systems on only the correct ones, predicting only the directionality of each rule for different values of β. Too low a value of β means that the algorithms tend to predict most things as unidirectional and too high a value means that the algorithms tend to predict everything as bidirectional. It is clear from the figure that the performance of all the systems reach their peak performance in the range $2 \leq \beta \leq 4$, which agrees with our intuition of obtaining the best system accuracy in a medium range. It is also seen that the best accuracy for each of the models goes up as compared to the corresponding values obtained in the general framework. The best performing system, IRM using CBC RSPs, reaches a peak accuracy of 63.64%, a much higher score than its accuracy score of 48% under the general framework and also a significant improvement over the baseline score of 48.48% for this task. Paired t-test shows that the difference is statistically significant at $p<0.05$. The baseline score for this task is obtained by assigning the most frequently occurring direction to all the correct inference rules. This paints a very encouraging picture about the ability of the algorithm to identify the directionality much more accurately if it can be provided with a cleaner set of inference rules.

6 Conclusion

Semantic inferences are fundamental to understanding natural language and are an integral part of many natural language applications such as question answering, summarization and textual entailment. Given the availability of large amounts of text and with the increase in computation power, learning them automatically from large text corpora has become increasingly feasible and popular. We introduced the Directionality Hypothesis, which states that if two paths share a significant number of relational selectional preferences (RSPs) and if the first has many more RSPs than the second, then the second path implies the first. Our experiments show empirical evidence that the Directionality Hypothesis with RSPs can indeed be used to filter incorrect inference rules and find the directionality of correct ones. We believe that this result is one step in the direction of solving the basic problem of semantic inference.

Several questions must still be addressed. The models need to be improved in order to address the problem of incorrect inference rules. The distributional hypothesis does not provide a framework to address the issue with antonymy relations like "*X loves Y*" ⇔ "*X hates Y*" and hence other ideas need to be investigated.

Ultimately, our goal is to improve the performance of NLP applications with better inferencing capabilities. Several recent data points, such as (Harabagiu and Hickl 2006), and others discussed in Section 2.1, give promise that refined inference rules for directionality may indeed improve question answering, textual entailment and multi-document summarization accuracies. It is our hope that methods such as the one proposed in this paper may one day be used to harness the richness of automatically created inference rule resources within large-scale NLP applications.

References

Anick, P.G. and Tipirneni, S. 1999. The Paraphrase Search Assistant: Terminology Feedback for Iterative Information Seeking. In *Proceedings of SIGIR 1999*. pp. 53-159. Berkeley, CA

Barzilay, R. and McKeown, K.R. 2001.Extracting Paraphrases from a Parallel Corpus. In *Proceedings of ACL 2001*. pp. 50–57. Toulose, France.

Barzilay, R.; McKeown, K.R. and Elhadad, M. 1999. Information Fusion in the Context of Multi-Document Summarization. In *Proceedings of ACL 1999*. College Park, Maryland.

Chklovski, T. and Pantel, P. 2004. VerbOCEAN: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of EMNLP 2004*. Barcellona Spain.

Cover, T.M. and Thomas, J.A. 1991. *Elements of Information Theory*. John Wiley & Sons.

Echihabi, A. and Marcu. D. 2003. A Noisy-Channel Approach to Question Answering. In *Proceedings of ACL 2003*. Sapporo, Japan.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Geffet, M.; Dagan, I. 2005. The Distributional Inclusion Hypothesis and Lexical Entailment. In *Proceedings of ACL 2005*. pp. 107-114. Ann Arbor, Michigan.

Harabagiu, S.; and Hickl, A. 2006. Methods for Using Textual Entailment in Open-Domain Question Answering. In *Proceedings of ACL 2006*. pp. 905-912. Sydney, Australia.

Harris, Z. 1954. Distributional structure. *Word*. 10(23): 146-162.

Lenat, D. 1995. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.

Lin, D. 1993. Parsing Without OverGeneration. In *Proceedings of ACL 1993*. pp. 112-120. Columbus, OH.

Lin, D. and Pantel, P. 2001. Discovery of Inference Rules for Question Answering. *Natural Language Engineering* 7(4):343-360.

Lin, D.; Zhao, S.; Qin, L. and Zhou, M. 2003. Identifying Synonyms among Distributionally Similar Words. In *Proceedings of IJCAI 2003*, pp. 1492-1493. Acapulco, Mexico.

Manning, C.D. and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

Moldovan, D.; Clark, C.; Harabagiu, S. and Maiorano S. 2003. COGEX: A Logic Prover for Question Answering. In *Proceedings of HLT/NAACL 2003*. Edmonton, Canada.

Pantel, P.; Bhagat, R.; Coppola, B.; Chklovski, T. and Hovy, E. 2007. ISP: Learning Inferential Selectional Preferences. In *Proceedings of HLT/NAACL 2007*. Rochester, NY.

Pantel, P. and Lin, D. 2002. Discovering Word Senses from Text. In *Proceedings of KDD 2002*. pp. 613-619. Edmonton, Canada.

Resnik, P. 1996. Selectional Constraints: An Information-Theoretic Model and its Computational Realization. *Cognition*, 61:127–159.

Siegel, S. and Castellan Jr., N. J. 1988. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill.

Szpektor, I.; Tanev, H.; Dagan, I.; and Coppola, B. 2004. Scaling web-based acquisition of entailment relations. In *Proceedings of EMNLP 2004*. pp. 41-48. Barcelona, Spain.

Torisawa, K. 2006. Acquiring Inference Rules with Temporal Constraints by Using Japanese Coordinated Sentences and Noun-Verb Co-occurances. In *Proceedings of HLT/NAACL 2006*. pp. 57-64. New York, New York.

Wilks, Y. 1975. Preference Semantics. In E.L. Keenan (ed.), *Formal Semantics of Natural Language*. Cambridge: Cambridge University Press.

Zanzotto, F.M.; Pennacchiotti, M.; Pazienza, M.T. 2006. Discovering Asymmetric Entailment Relations between Verbs using Selectional Preferences. In *Proceedings of COLING/ACL 2006*. pp. 849-856. Sydney, Australia.