Using Relational Selectional Preferences to Improve Inference Resources

Learning Directionality and Selectional Preferences of Inference Rules

Miriam Käshammer

Course: Recent Developments in Computational Semantics
Universität des Saarlandes

May 31, 2010

Recap

Inference rule/paraphrase collections are ...

- known to improve performance of various NLP tasks (e.g. IR, QA, Summarization)
- automatically built from text
  - corpus: DIRT [Lin and Pantel, 2001]
  - web: TE/ASE [Szpektor et al., 2004]

Example

`X writes Y` ⇔ `X is the author of Y`

Downside of automatic approaches I

Inference rules are underspecified in directionality

`X eats Y` $\Leftrightarrow$ `X likes Y` (DIRT)

$\underline{\textit{He}}$ eats $\underline{\textit{spicy food}} \Rightarrow \underline{\textit{He}}$ likes $\underline{\textit{spicy food}}$
$\underline{\textit{He}}$ eats $\underline{\textit{rollerblading}} \not\Leftarrow \underline{\textit{He}}$ likes $\underline{\textit{rollerblading}}$

`X eats Y` $\Rightarrow$ `X likes Y`

Downside of automatic approaches II

Blind application of inference rules, regardless of context or word senses

`X is charged by Y` ⇒ `Y announced the arrest of X`

Nichols `was charged by` *federal prosecutors for murder*
⇒ *Federal prosecutors* `announced the arrest of` *Nichols*

Accounts `were charged by` *CCM telemarketers without obtaining authorizations*
⇏ *CCM telemarketers* `announced the arrest of` *accounts*

Towards a solution

LEDIR: An Unsupervised Algorithm for Learning Directionality of Inference Rules.
Bhagat, R., Pantel, P., and Hovy, E. (2007).

Goal: Identify the directionality of inference rules.

ISP: Learning Inferential Selectional Preferences.
Pantel, P., Bhagat, R., Coppola, B., Chklovski, T., and Hovy, E. (2007).

Goal: Learn admissible argument values to which an inference rule can be applied.

⇒ Relational Selectional Preferences

Outline

1. Introduction
2. Relational Selectional Preferences
3. Learning Directionality of Inference Rules
4. Learning Selectional Preferences for Inference Rules
5. Conclusions

Outline

1. Introduction
2. Relational Selectional Preferences
   - Background
   - Learning RSP
3. Learning Directionality of Inference Rules
4. Learning Selectional Preferences for Inference Rules
5. Conclusions

Relations and selectional preferences

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 $\langle x, p, y \rangle$ be an instance of the relation p.

Relational selectional preferences (RSP) of a binary relation p

The set of semantic classes C_x and C_y of the words that can occur in positions x and y respectively.

Example

$p =$ `X likes Y`
RSP for `X`: $C_x = \{$`Individual`, `Social_Group`, ...$\}$
RSP for `Y`: $C_y = \{$`Individual`, `Food`, `Activity`, ...$\}$

Semantic classes

⇒ Choice of semantic classes (e.g. granularity) is crucial for learning RSP

⇒ No ideal set of universally acceptable semantic classes available

- Manually created taxonomy (e.g. WordNet)
- Automatically generated classes from the output of a word clustering algorithm

Learning RSP

$p =$ `X is charged by Y`

Joint selectional preferences

$\langle$ Person, p, Law_Inforcement_Agent$\rangle$

$\langle$ Person, p, Law_Inforcement_Agency$\rangle$

$\langle$ Bank_Account, p, Organization$\rangle$ …

Independent selectional preferences

$\langle$ Person, p, *$\rangle$

$\langle$ *, p, Law_Inforcement_Agency$\rangle$

$\langle$ *, p, Organization$\rangle$ …

⇒**Two models for learning RSP based on corpus analysis**

Joint Relational Model (JRM)

Obtaining candidates

Given a relation p and a large corpus of (English) text:

1. Find all occurrences of p.
2. For each instance $\langle x, p, y \rangle$:
   - Obtain the sets C_x and C_y of semantic classes that x and y belong to.
   - Every triple $\langle c_x, p, c_y \rangle$ is a **candidate selectional preference** for p, by assuming that every $c_x \in C_x$ can co-occur with every $c_y \in C_y$ and vice versa.

The set of RSPs for p: $\langle C_x, p, C_y \rangle$

Joint Relational Model (JRM)

Ranking candidates I

A candidate can be incorrect when

- it was generated from the incorrect sense of a polysemous word, or
- p does not hold for the other words in the semantic class.

We have more confidence in a particular candidate if its semantic classes are closely related given the relation p:

Pointwise mutual information

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

Joint Relational Model (JRM)

Ranking candidates II

Maximum likelihood estimates over the corpus

$$P(c_x|p) = \frac{|c_x, p, *|}{|*, p, *|} \quad P(c_y|p) = \frac{|*, p, c_y|}{|*, p, *|} \quad P(c_x, c_y|p) = \frac{|c_x, p, c_y|}{|*, p, *|}$$

$|c_x, p, c_y|$: frequency of observing $\langle c_x, p, c_y \rangle$

$$|c_x, p, *| = \sum_{w \in c_x} \frac{|w, p, *|}{|C(w)|} \qquad |*, p, c_y| = \sum_{w \in c_y} \frac{|*, p, w|}{|C(w)|}$$

$$|c_x, p, c_y| = \sum_{w_1 \in c_x, w_2 \in c_y} \frac{|w_1, p, w_2|}{|C(w_1)| \cdot |C(w_2)|}$$

$|x, p, y|$: frequency of observing $\langle x, p, y \rangle$

$|C(w)|$: number of classes to which w belongs

Independent Relational Model (IRM)

Obtaining and ranking candidates

Given a relation p and a large corpus of (English) text:

1. Find all occurrences of p.
2. For each instance $\langle x, p, y \rangle$:
   - Obtain the sets C_x and C_y of semantic classes that x and y belong to.
   - All triples $\langle c_x, p, * \rangle$ and $\langle *, p, c_y \rangle$ are **candidate selectional preferences** for p, where $c_x \in C_x$ and $c_y \in C_y$.

Use MLE for $P(c_x|p)$ and $P(c_y|p)$ to rank the candidates.

Independent Relational Model (IRM)

Joint representation

Joint representation of independently learnt RSPs

Cartesian product of the sets $\langle C_x, p, *\rangle$ and $\langle *, p, C_y\rangle$

$$\langle C_x, p, *\rangle \times \langle *, p, C_y\rangle = \left\{ \begin{array}{ll} \langle c_x, p, c_y\rangle : & \langle c_x, p, *\rangle \in \langle C_x, p, *\rangle \text{ and} \\ & \langle *, p, c_y\rangle \in \langle *, p, C_y\rangle \end{array} \right\}$$

Outline

1. Introduction
2. Relational Selectional Preferences
3. Learning Directionality of Inference Rules
    - Algorithm
    - Evaluation
4. Learning Selectional Preferences for Inference Rules
5. Conclusions

Problem definition

Goal: Filter out incorrect inference rules and identify the directionality of the correct ones.

Formally

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*

Underlying assumption

Distributional Hypothesis:

Words that appear in the same contexts tend to have similar meanings. (Harris, 1954)

Underlying assumption

Distributional Hypothesis:

Words that appear in the same contexts tend to have similar meanings. (Harris, 1954)

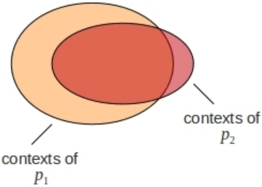
Extension: Directionality Hypothesis

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 one most likely implies the first and not vice versa.

Underlying assumption

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 one most likely implies the first and not vice versa.



$$p_1 \Leftarrow p_2$$

$$\texttt{X likes Y} \Leftarrow \texttt{X eats Y}$$

Steps of the algorithm

Given a candidate inference rule $p_i \Leftrightarrow p_j$:

1. Model the contexts of p_i and p_j by selectional preferences (RSP).
2. Determine the plausability of the inference rule.
3. If it is plausible, determine its directionality.

Inference plausability

Overlap coefficient between two sets

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

Overlap coefficient between the RSP of p_i and p_j

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

Given a candidate inference rule $p_i \Leftrightarrow p_j$ and the respective RSPs:

If $sim(p_i, p_j) \geq \alpha$ $\rightarrow$ *inference is plausible*
else $\rightarrow$ *inference is not plausible*

Directionality model

For a plausible inference:

$$\begin{array}{llll}
\text{If} & \dfrac{|C_x, p_i, C_y|}{|C_x, p_j, C_y|} \geq \beta & \rightarrow & p_i \Leftarrow p_j \\
\text{else if} & \dfrac{|C_x, p_i, C_y|}{|C_x, p_j, C_y|} \leq \dfrac{1}{\beta} & \rightarrow & p_i \Rightarrow p_j \\
\text{else} & & \rightarrow & p_i \Leftrightarrow p_j
\end{array}$$

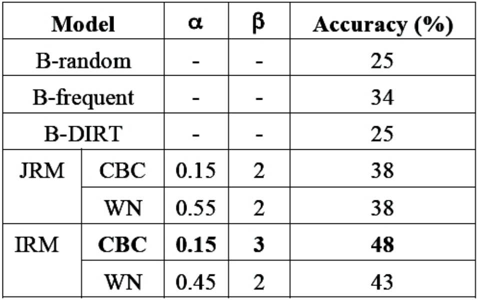
with $\beta \geq 1$

Experimental Setup

- Two sets of **semantic classes**:
  - WordNet (WN) synsets at depth four: 1287 semantic classes
  - Clustering algorithm (CBC) (Pantel and Lin, 2002) on newswire collections: 1628 semantic classes
- 1999 AP newswire collection (31 million words), Minipar parser
- Manually annotated **gold standard**:
  - four tags: $\Leftrightarrow$ / $\Rightarrow$ / $\Leftarrow$ / NO
  - Development set: 57 DIRT inference rules
  - Test set: 100 DIRT inference rules
- Development phase: Experiments with different parameter (α, β) combinations on the development set to obtain the best performing parameter combination for each of the four system.
- Evaluation criterion: Accuracy $= \frac{\#\text{ correctly tagged inferences}}{\#\text{ all input inferences}}$

Results I

Results on the test set

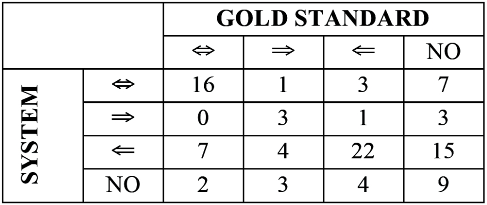


Model		α	β	Accuracy (%)
B-random		-	-	25
B-frequent		-	-	34
B-DIRT		-	-	25
JRM	CBC	0.15	2	38
	WN	0.55	2	38
IRM	**CBC**	**0.15**	**3**	**48**
	WN	0.45	2	43

B-random: Randomly assigns one of the four tags to each rule.
B-frequent: Assigns the most frequent tag in the gold standard to each rule.
B-DIRT: Assigns the bidirectional tag to each rule.

Results II

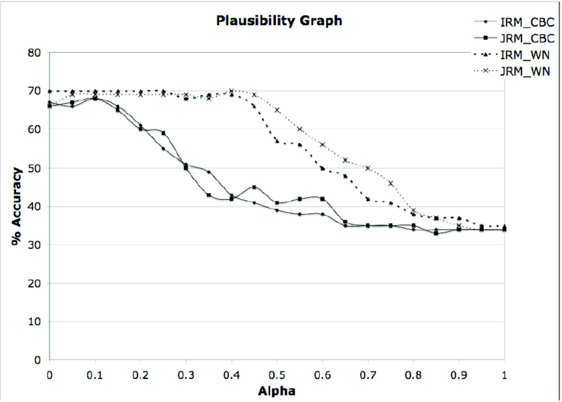
Confusion matrix for the best performing system:
IRM using CBC with $\alpha = 0.15$ and $\beta = 3$



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

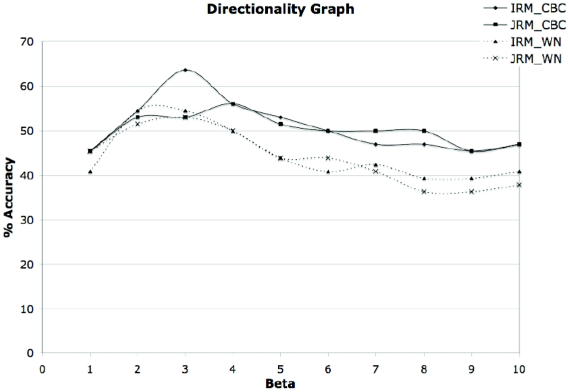
Results III

Accuracy variation in predicting correct vs. incorrect inference rules for different values of α



Results IV

Accuracy variation in predicting directionality of correct inference rules for different values of β



Outline

1. Introduction
2. Relational Selectional Preferences
3. Learning Directionality of Inference Rules
4. Learning Selectional Preferences for Inference Rules
   - Algorithm
   - Evaluation
5. Conclusions

Problem definition

Goal: Learn inferential selectional preferences for filtering inference rules.

Formally

Given an inference rule $p_i \Rightarrow p_j$ and the instance $\langle x, p_i, y \rangle$, determine if $\langle x, p_j, y \rangle$ is valid.

Example

`X is charged by Y` $\Rightarrow$ `Y announced the arrest of X`

Accounts were charged by CCM telemarketers without obtaining authorizations

$\overset{??}{\Rightarrow}$ *CCM telemarketers announced the arrest of accounts*

Inferential selectional preferences

Obtaining and ranking candidates

Inferential selectional preferences (ISP) for $p_i \Rightarrow p_j$

The intersection of the relational selectional preferences (RSP) for p_i and p_j

Ways to **rank** the candidates by combining their RSP scores:

- minimum
- maximum
- average

Inferential selectional preferences

Joint Inferential Model (JIM)

Example

$p_i =$ `X is charged by Y`

$$\left.\begin{array}{l}\langle\ \texttt{Person},\ p_i,\ \texttt{Law\_Enforcement\_Agent}\ \rangle = 1.45\\ \langle\ \texttt{Person},\ p_i,\ \texttt{Law\_Enforcement\_Agency}\ \rangle = 1.21\\ \langle\ \texttt{Bank\_Account},\ p_i,\ \texttt{Organization}\ \rangle = 0.97\end{array}\right\}\text{RSP}$$

$p_j =$ `Y announced the arrest of X`

$$\left.\begin{array}{l}\langle\ \texttt{Person},\ p_j,\ \texttt{Law\_Enforcement\_Agent}\ \rangle = 2.01\\ \langle\ \texttt{Person},\ p_j,\ \texttt{Reporter}\ \rangle = 1.98\\ \langle\ \texttt{Person},\ p_j,\ \texttt{Law\_Enforcement\_Agency}\ \rangle = 1.61\end{array}\right\}\text{RSP}$$

$p_i \Rightarrow p_j$

$$\left.\begin{array}{l}\langle\ \texttt{Person},\ \texttt{Law\_Enforcement\_Agent}\ \rangle = 1.45/2.01/1.73\\ \langle\ \texttt{Person},\ \texttt{Law\_Enforcement\_Agency}\ \rangle = 1.21/1.61/1.41\end{array}\right\}\text{ISP}$$

Inferential selectional preferences

Independent Inferential Model (IIM)

Example

$$p_i = \texttt{X is charged by Y}$$

$$\left.\begin{array}{l} \langle\ \texttt{*},\ p_i,\ \texttt{Law\_Enforcement\_Agent}\ \rangle = 3.43 \\ \langle\ \texttt{Person},\ p_i,\ \texttt{*}\ \rangle = 2.17 \\ \langle\ \texttt{*},\ p_i,\ \texttt{Organization}\ \rangle = 1.24 \end{array}\right\}\text{RSP}$$

$$p_j = \texttt{Y announced the arrest of X}$$

$$\left.\begin{array}{l} \langle\ \texttt{Person},\ p_j,\ \texttt{*}\ \rangle = 2.87 \\ \langle\ \texttt{*},\ p_j,\ \texttt{Law\_Enforcement\_Agent}\ \rangle = 1.92 \\ \langle\ \texttt{*},\ p_j,\ \texttt{Reporter}\ \rangle = 0.89 \end{array}\right\}\text{RSP}$$

$$p_i \Rightarrow p_j$$

$$\left.\begin{array}{l} \langle\ \texttt{*},\ \texttt{Law\_Enforcement\_Agent}\ \rangle = 1.92/3.43/2.675 \\ \langle\ \texttt{Person},\ \texttt{*}\ \rangle = 2.17/2.87/2.52 \end{array}\right\}\text{ISP}$$

Filtering Algorithms

Given $p_i \Rightarrow p_j$ and $\langle x, p_i, y\rangle$, three different algorithms are proposed to check whether $\langle x, p_j, y\rangle$ is valid:

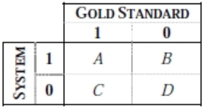
- **ISP.JIM**:
  The ISP $\langle c_x, c_y\rangle$ (for some $c_x \in C_x$ and $c_y \in C_y$) was admitted by the Joint Inferential Model.
- **ISP.IIM.AND**:
  The ISPs $\langle c_x, *\rangle$ and $\langle *, c_y\rangle$ (for some $c_x \in C_x$ and $c_y \in C_y$) were admitted by the Independent Inferential Model.
- **ISP.IIM.OR**:
  The ISP $\langle c_x, *\rangle$ or $\langle *, c_y\rangle$ (for some $c_x \in C_x$ and $c_y \in C_y$) was admitted by the Independent Inferential Model.

Furthermore: Select only the top τ percent hightest ranking ISPs.

Experimental Setup

Similar setup as in the previous section:

- Two sets of **semantic classes**:
  - WordNet (WN) synsets at depth four: 1287 semantic classes
  - Clustering algorithm (CBC) (Pantel and Lin, 2002) on newswire collections: 1628 semantic classes
- 1999 AP newswire collection (31 million words), Minipar parser
- **Gold standard** construction:
  - 100 DIRT inference rules $p_i \Rightarrow p_j$
  - 10 randomly selected instances per p_i: $\langle x, p_i, y \rangle$
  - Question: Is $\langle x, p_j, y \rangle$ valid and does the inference hold?
  - Development and test set, 500 instances $\langle x, p_j, y \rangle$ each
- Evaluation criteria:
  - Sensitivity $= \frac{A}{A+C}$
  - Specificity $= \frac{D}{B+D}$
  - Accuracy $= \frac{A+D}{A+B+C+D}$



		GOLD STANDARD	
		1	0
SYSTEM	1	A	B
	0	C	D

Development Phase

- Experiments on the development set with different parameter combinations (ranking strategy, τ) for each of the six system
- Select the best parameter combination according to:
  - ***Accuracy***: Overall ability to correctly accept and reject inferences
  - ***90%-Specificity***: Best sensitivity while maintaining at least 90% specificity
- Evaluation of the selected systems on the test set

Results I

Best performing systems

(Selection based on the *Accuracy* criterion)

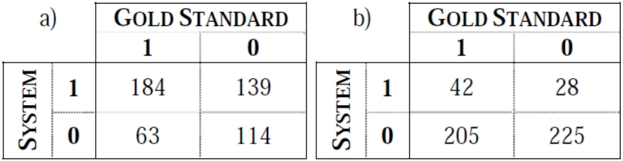
System		*Ranking*	τ *(%)*	*Sensit.*	*Specif.*	*Acc.*
B0		-	-	0.00	1.00	0.50
B1		-	-	1.00	0.00	0.49
Random		-	-	0.50	0.47	0.50
CBC	**ISP.JIM**	**max**	**100**	**0.17**	**0.88**	**0.53**$^{+}$
	ISP.IIM.AND	max	100	0.24	0.84	0.54
	ISP.IIM.OR	**max**	**90**	**0.73**	**0.45**	**0.59***
WN	ISP.JIM	min	40	0.20	0.75	0.47
	ISP.IIM.AND	min	10	0.33	0.77	0.55
	ISP.IIM.OR	min	20	0.87	0.17	0.51

* significantly better than the three baselines

$^{+}$ best system according to the *90%-Specificity* criterion

Results II

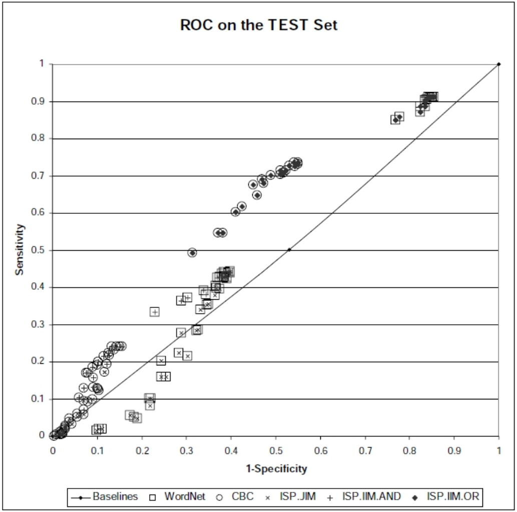
Confusion matrices for
a) ISP.IIM.OR - best *Accuracy*
b) ISP.JIM - best *90%-Specificity*



a)		GOLD STANDARD	
		1	0
SYSTEM	1	184	139
	0	63	114

b)		GOLD STANDARD	
		1	0
SYSTEM	1	42	28
	0	205	225

Results III


ROC on the TEST Set
Sensitivity
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1-Specificity
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Baselines
□ WordNet
○ CBC
× ISP.JIM
+ ISP.IIM.AND
◆ ISP.IIM.OR

Conclusions

- Empirical evidence that **relational selectional preferences** ...
  - and the Directionality Hypothesis can be used to determine the plausability and **directionality** of inference rules.
  - can be used to learn **admissible argument values** for inference rules.
- More research regarding the appropriate **inventory of semantic classes** for selectional preferences is necessary.
- Additional models for filtering incorrect rules are needed (problem of antonymy).

References

Bhagat, R., Pantel, P., and Hovy, E. (2007).
LEDIR: An Unsupervised Algorithm for Learning Directionality of Inference Rules.
Proceedings of EMNLP-CoNLL 2007.

Lin, D. and Pantel, P. (2001).
DIRT - Discovery of Inference Rules from Text.
Proceedings of ACM Conference on Knowledge Discovery and Data Mining.

Pantel, P., Bhagat, R., Coppola, B., Chklovski, T., and Hovy, E. (2007).
ISP: Learning Inferential Selectional Preferences.
Proceedings of NAACL/HLT 2007.

Szpektor, I., Tanev, H., Dagan, I., and Coppola, B. (2004).
Scaling Web-based Acquisition of Entailment Relations.
Proceedings of EMNLP 2004.