Learning genic interactions without expert domain knowledge

Luboš Popelínský and Jan Blaťák
Knowledge Discovery Lab, Faculty of Informatics
Masaryk University in Brno

no expert knowledge in the GE domain available

combining positive and disambiguation rules
simple and complex interactions solved separately

NLP tools; first-order frequent patterns as new features

data without coreference

Two-step learning

first step – learning rules for sentences that **contain a single pair of terms**

second step – learning rules for **all sentences**

In each step

positive rules – for a given sentence return a pair of agent-target, and

disambiguation rules – from all possible pairs of tags (agent, target) an incorrect pair is removed – are learned

positive rules are applied first

disambiguation rules remove the remaining ambiguities

Domain knowledge

added

`POS tags` – Brill tagger

`hyperonyma` – WordNet

`ffverb` – returns a verb that has appeared between two terms (agents, targets)

removed

`lemma` – it has almost never appeared in the learned rules

`word` – resulted in speed up of learning without accuracy decrease

Learning tools

RAP

- *frequent syntactic patterns* – `relation`, `ffverb` and `follows(Word1,Word2)`
- min. support 10%, max. length 15 literals
- best-first search, entropy based heuristics that prefers emerging patterns
- learning class association rules

Aleph

- learning positive and disambiguation rules
- with or without the frequent syntactic patterns
- `clauselength=5`

Weka

- All 536 patterns with non-zero support, found with RAP
- SVM, J4.8, Naive Bayes classifier, instance-based learner IB1

Algorithm

Given POSRULES, MINPOS, DISRULES and MINNEG

A1 and A2 = valid genic interaction pair (Agent,Target), if

Apply positive rules

(i) at least POSRULES rules have fired, or

(ii) a single rule has fired that covered at least MINPOS positive examples from the learning set, and

(iii) there is no (A2,A1) after application of all the positive rules.

Apply disambiguation rules

(i) at least DISRULES rules have fired, or

(ii) a single rule has fired that covered at least MINNEG negative examples from the learning set.

Summary of results

		PRE	REC	F-M
AL2	**Aleph, 2-step method**	**46.5**	**50.0**	**48.2**
AFP	Aleph + freq.patterns	37.6	64.8	47.6
AL1	Aleph, no freq.patterns	42.5	42.5	42.5
CAR	class association rules	37.2	29.6	32.9
PRO	propositionalization	28.0	29.6	28.8
LLL	Aleph, 2-step method	**37.9**	**55.5**	**45.1**

Two-step learning: top 5 results

MINPOS	POSRUL	MINNEG	DISRUL	F-M
5	3	3	2	48.2
6	3	3	2	46.7
5	3	2	2	45.7
5	3	0	0	45.6
4	2	0	0	45.1

Single-step learning

MINPOS	POSRUL	MINNEG	DISRUL	F-M
4	2	3	2	42.5
3	2	0	3	41.8
3	2	3	2	41.5
3	2	2	1	40.0
3	2	0	0	38.0

Single-step learning: Maximizing precision

MINPOS	POSRUL	COR	PRE	REC	F-M
6	5	17	62.9	31.4	41.9
6	4	17	60.7	30.6	40.1
7	4	17		dtto	
7	3	18	60.0	33.3	42.8

Weka

Results with propositionalized data

	PRE	REC	F-M
SVM	28.0	29.6	28.8
Decision tree	35.4	20.3	25.8
Naive Bayes	22.5	16.6	19.1
IB1	16.4	22.2	18.8

features = all 536 patterns found with RAP, with non-zero support

Discussion

Frequent patterns

5.1% increase of F-measure

but was not higher then the best result (two-step learning)

appearance of the patterns in the rules – 10%.

Domain knowledge

without POS tagging with Brill tagger – 10% decrease of F-measure

without `hyper` – much smaller effect

appearance in the learned rules:
`tag` in 32.4% rules, `ffverb` 34.3%, `hyper` 15.6%