Generic Ontology Learners on Application Domains

Francesca Fallucchi[1] Maria Teresa Pazienza [1]
Fabio Massimo Zanzotto[1]

[1]DISP
University of Rome Tor Vergata
Rome, Italy
`{fallucchi,pazienza,zanzotto}@info.uniroma2.it`

LREC 2010, Malta, May 2010

Motivation

- Learning methods require large general corpora and knowledge repositories
- In specific domains ontologies are extremely poor
- Manually building ontologies is a very time consuming and expensive task
- Automatically creating or extending ontologies needs large corpora and existing structured knowledge to achieve reasonable performance

Motivation

Problems

- Scarcity of domains covered by existing ontologies
- Not relevant existing ontologies to expand for target domain

Motivation

Problems

- Scarcity of domains covered by existing ontologies
- Not relevant existing ontologies to expand for target domain

$\Downarrow$

Solution

- We propose a model that can be used in different specific knowledge domains with a small effort for its adaptation
- Our model is learned from a generic domain that can be exploited to extract new informations in a specific domain

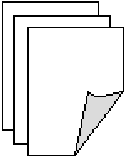
Our Learner Model

- Model exploits the information learned in a background domain for extracting information in an adaptation domain
- Model is based on the probabilistic formulation
- Model takes into consideration corpus-extracted evidences over a list of training pairs
- Model is used to estimate the probabilities of the new instances computing a new feature space

Corpus Analysis

Corpus Analysis

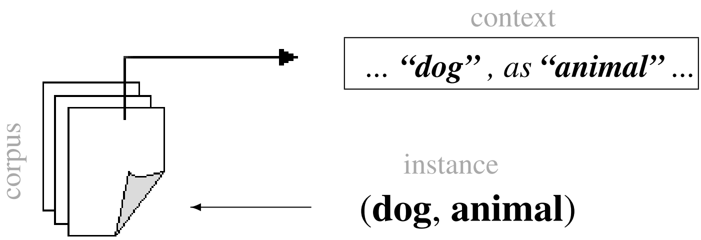
corpus



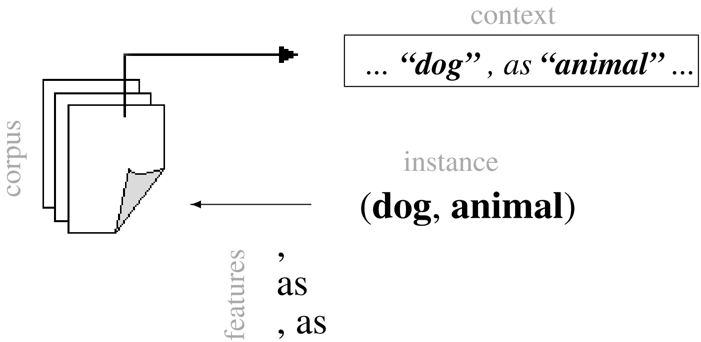
instance

(dog, animal)

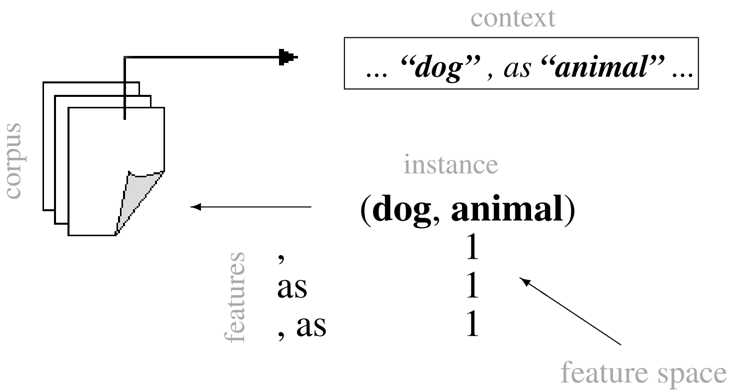
Corpus Analysis



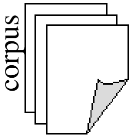
Corpus Analysis



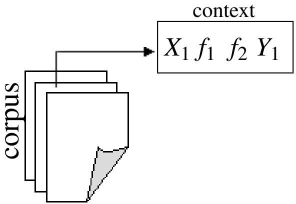
Corpus Analysis



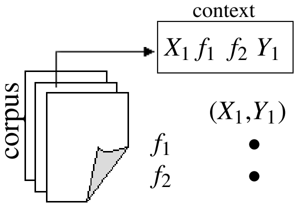
Corpus Analysis



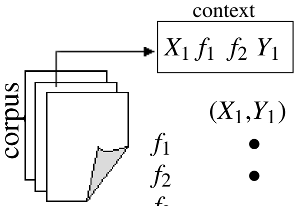
Corpus Analysis



Corpus Analysis

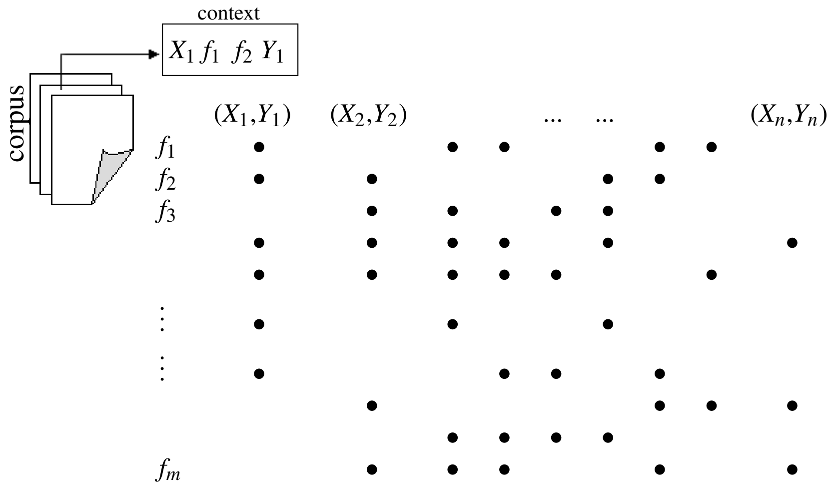


Corpus Analysis

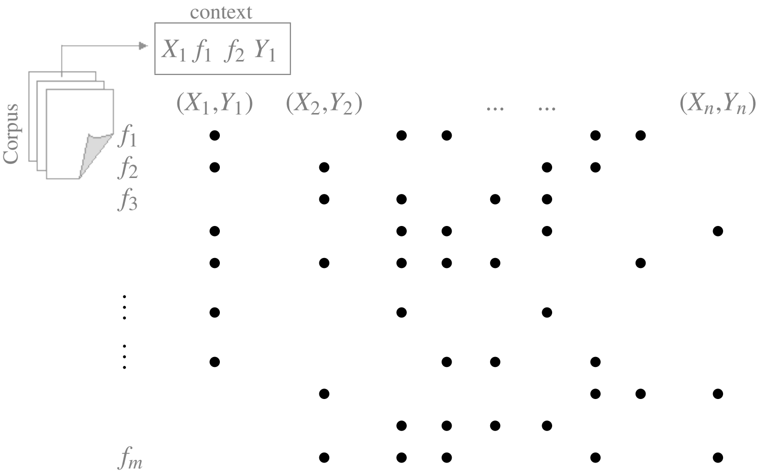


	(X_1, Y_1)	(X_2, Y_2)
f_1	•	
f_2	•	•
f_3		•

Corpus Analysis



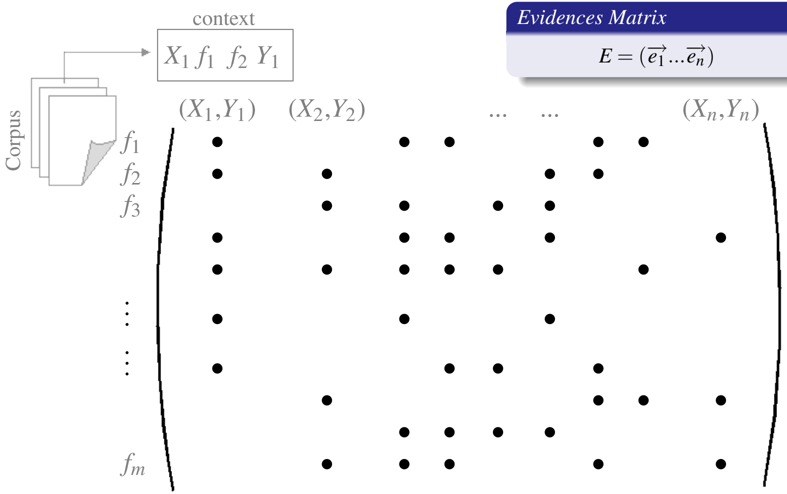
Instances Matrix



Instances Matrix

Evidences Matrix

$$E = (\overrightarrow{e_1} \ldots \overrightarrow{e_n})$$



A Probabilistic Model

Probabilistic model for learning ontologies form corpora

- Ontology is seen as a set O of relations R over pairs $R_{i,j}$
- If $R_{i,j}$ is in O, i is a concept and j is one of its generalization

Goal: Estimate Posterior Probability

$$P(R_{i,j} \in O|E)$$

where E is a set of evidences extracted from corpus

Logistic Regression

Logit

Given two variables Y and X, the probability p of Y to be 1 given that $X = x$ is: $p = P(Y = 1|X = x)$ and $Y \sim Bernoulli(p)$

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Logistic Regression

Logit

Given two variables Y and X, the probability p of Y to be 1 given that $X = x$ is: $p = P(Y = 1|X = x)$ and $Y \sim Bernoulli(p)$

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

$$logit(p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Logistic Regression

Logit

Given two variables Y and X, the probability p of Y to be 1 given that $X = x$ is: $p = P(Y = 1|X = x)$ and $Y \sim Bernoulli(p)$

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

$$logit(p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

Given regression coefficients the probability is

$$p(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k)}$$

Estimating Regression Coefficients

We estimate the regressors $\beta_0, \beta_1, ..., \beta_k$ of $x_1, ..., x_k$ with

- maximal likelihood estimation
- $logit(p) = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k$
- solving a linear problem

Estimating Regression Coefficients

We estimate the regressors $\beta_0, \beta_1, \ldots, \beta_k$ of $x_1, \ldots, x_k$ with

- maximal likelihood estimation
- $logit(p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$
- solving a linear problem

$$\overrightarrow{logit(p)} = E\beta$$

where

$$E = \begin{pmatrix} 1 & e_{11} & e_{12} & \cdots & e_{1n} \\ 1 & e_{21} & e_{22} & \cdots & e_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & e_{m1} & e_{m2} & \cdots & e_{mn} \end{pmatrix}$$

Background Ontology Learner

Using a logistic regressor based on the Moore-Penrose pseudo-inverse matrix (Fallucchi and Zanzotto, RANLP 2009)

$$\widehat{\beta} = X^{+}_{C_B} l$$

where:

- $X^{+}_{C_B}$ is the pseudo-inverse matrix of the evidences matrix X_{C_B} obtained from a generic corpus C_B
- l is the logit vector $(\overrightarrow{logit(p)})$

Estimator for Application Domain

The logit of the testing pairs

$$l' = \alpha X_{C_A} \widehat{\beta}$$

where:

- α is a parameter used to adapt the model by the β vector to the new domain
- X_{C_A} is the inverse evidence matrix obtained from an *adaptation* domain corpus C_A
- $\widehat{\beta}$ is the regressors vector

Estimator for Application Domain

The logit of the testing pairs

$$l' = \alpha X_{C_A} \widehat{\beta}$$

where:

- α is a parameter used to adapt the model by the β vector to the new domain
- X_{C_A} is the inverse evidence matrix obtained from an *adaptation* domain corpus C_A
- $\widehat{\beta}$ is the regressors vector

Then, step by step testing pairs probability

$$p_i = \frac{\exp(l_i)}{1+\exp(l_i)}$$

Artificial Intelligence at Roma Tor Vergata

1. Motivations
2. Probabilistic Ontology Learning
   - Corpus Analysis
   - A Probabilistic Model
   - Logistic Regression
3. Experimental Evaluation
   - Experimental Set-Up
   - Agreement
   - Results
4. Conclusions and Future Works

Experimental Set-Up

1. *Target Ontologies*
   - Training: pairs that are in hyperonym relation in WordNet ==> about 600000 pairs of words
   - Testing: pairs in Earth Observation Domain ==> about 404 pairs of words
2. *Corpus*
   - Training: *English Web as Corpus*, ukWaC (Ferraresi,2008) ==> about 2700000 web pages
   - Testing: corpus related to *Earth Observation Domain* ==> about 8300 web pages
3. *Feature Spaces*
   - bag-of-words and n-grams
   - windows: length 3 tokens
   - ==> about 280000 features

Annotators for Testing Pairs

- Three annotators (A_1, A_2 and A_3) to build three different ontologies
- Two annotators are expert in the domain (A_1 and A_2), the third one is not (A_3)
- A_1 and A_2 have different levels of expertise: A_1 is a young expert in the domain and A_2 an older one
- Each annotator made a binary classification of 641 pairs of words in Earth Observation Domain

Annotators for Testing Pairs

- Three annotators (A_1, A_2 and A_3) to build three different ontologies
- Two annotators are expert in the domain (A_1 and A_2), the third one is not (A_3)
- A_1 and A_2 have different levels of expertise: A_1 is a young expert in the domain and A_2 an older one
- Each annotator made a binary classification of 641 pairs of words in Earth Observation Domain
  Only 404 pairs are found in *Earth Observation Corpus*

Evaluating the Quality of Annotations

Quality of the annotation procedure according to inter-annotation agreement among annotators

- **Pairwise Agreement**
  - Inter-annotators agreement for each pair of annotators
  - Contigency table
- **Multi$-\pi$ Agreement**
  - Inter-annotators agreement for all annotators together
  - Agreement table

Pairwise Agreement 404-annotation

		A_1 yes	A_1 no	
A_2	yes	40	32	72
	no	35	297	332
		75	329	404

$pair_1 = (A_1, A_2)$

		A_1 yes	A_1 no	
A_3	yes	65	54	119
	no	10	275	285
		75	329	404

$pair_2 = (A_1, A_3)$

		A_2 yes	A_2 no	
A_3	yes	53	66	119
	no	19	266	285
		72	332	404

$pair_3 = (A_2, A_3)$

Table: Contingency tables for pairwise annotator agreement

Pairwise Agreement 404-annotation

		A_1 yes	A_1 no	
A_2	yes	40	32	72
	no	35	297	332
		75	329	404

$pair_1 = (A_1, A_2)$

		A_1 yes	A_1 no	
A_3	yes	65	54	119
	no	10	275	285
		75	329	404

$pair_2 = (A_1, A_3)$

		A_2 yes	A_2 no	
A_3	yes	53	66	119
	no	19	266	285
		72	332	404

$pair_3 = (A_2, A_3)$

Table: Contingency tables for pairwise annotator agreement

	A_o	A_e	*kappa*
$pair_1 = (A_1, A_2)$	0.8341584	0.7023086	0.4429077
$pair_2 = (A_1, A_3)$	0.8415842	0.6291663	0.5728117
$pair_3 = (A_2, A_3)$	0.7896040	0.6322174	0.4279336

Table: pairwise agreement

Multi−π Agreement 404-annotation

pairs of words	A_1	A_2	A_3	Yes	No
(forest,terra firma)	1	1	1	3	0
(wind,process)	0	0	0	0	3
(forest,object)	0	0	0	0	3
(cloud,state)	0	1	0	1	2
(soil,object)	0	1	1	2	1
(wind,breath)	0	0	0	0	3
(wind,act)	0	0	0	0	3
(topography,geography)	1	1	1	3	0
...	...	...	...	...	...
TOTAL	75	72	119	266 (0.22)	946 (0.78)

Table: Agreement table

Multi—π Agreement 404-annotation

pairs of words	A_1	A_2	A_3	Yes	No
(forest,terra firma)	1	1	1	3	0
(wind,process)	0	0	0	0	3
(forest,object)	0	0	0	0	3
(cloud,state)	0	1	0	1	2
(soil,object)	0	1	1	2	1
(wind,breath)	0	0	0	0	3
(wind,act)	0	0	0	0	3
(topography,geography)	1	1	1	3	0
...	...	...	...	...	...
TOTAL	75	72	119	266 (0.22)	946 (0.78)

Table: Agreement table

Multi-π agreement

$$A_o = 0.82382 \qquad A_e = 0.65739$$

$$kappa = 0.48577$$

Experiments

Objective

To compute a model using both a *background* domain and an existing ontology can be positively used to learn the *isa* relation in Earth Observation Domain.

Experiments

Objective

To compute a model using both a *background* domain and an existing ontology can be positively used to learn the *isa* relation in Earth Observation Domain.

We compare two systems

- *WN-System*: existing hyperonym links in WordNet
- *Our-System*: our learner model

measuring their performance to replicate the three target ontologies produced by the three annotators

Results

annotators	recall	precision	f-measure
A_1	0,36	0.184932	0,244344
A_2	0,305556	0,150685	0,201836
A_3	0,470588	0,383562	0,422642

Table: *WN-System* against the 3 annotators

Results

annotators	recall	precision	f-measure
A_1	0,36	0.184932	0,244344
A_2	0,305556	0,150685	0,201836
A_3	0,470588	0,383562	0,422642

Table: *WN-System* against the 3 annotators

annotators	recall	precision	f-measure
A_1	0,493333	0,253425	0,334842
A_2	0,4305556	0,212329	0,284404
A_3	0,4369748	0,356164	0,392453

Table: *Our-System* against the 3 annotators

Conclusions

- We propose a model adaptation strategy that use a *background* domain to learn the *isa* relations in a specific domain
- Experiments show that this way of using a model identified in a *background* domain is helpful to learn the *isa* relation in Earth Observation Domain.
- We will try to learn ontologies in other target domain