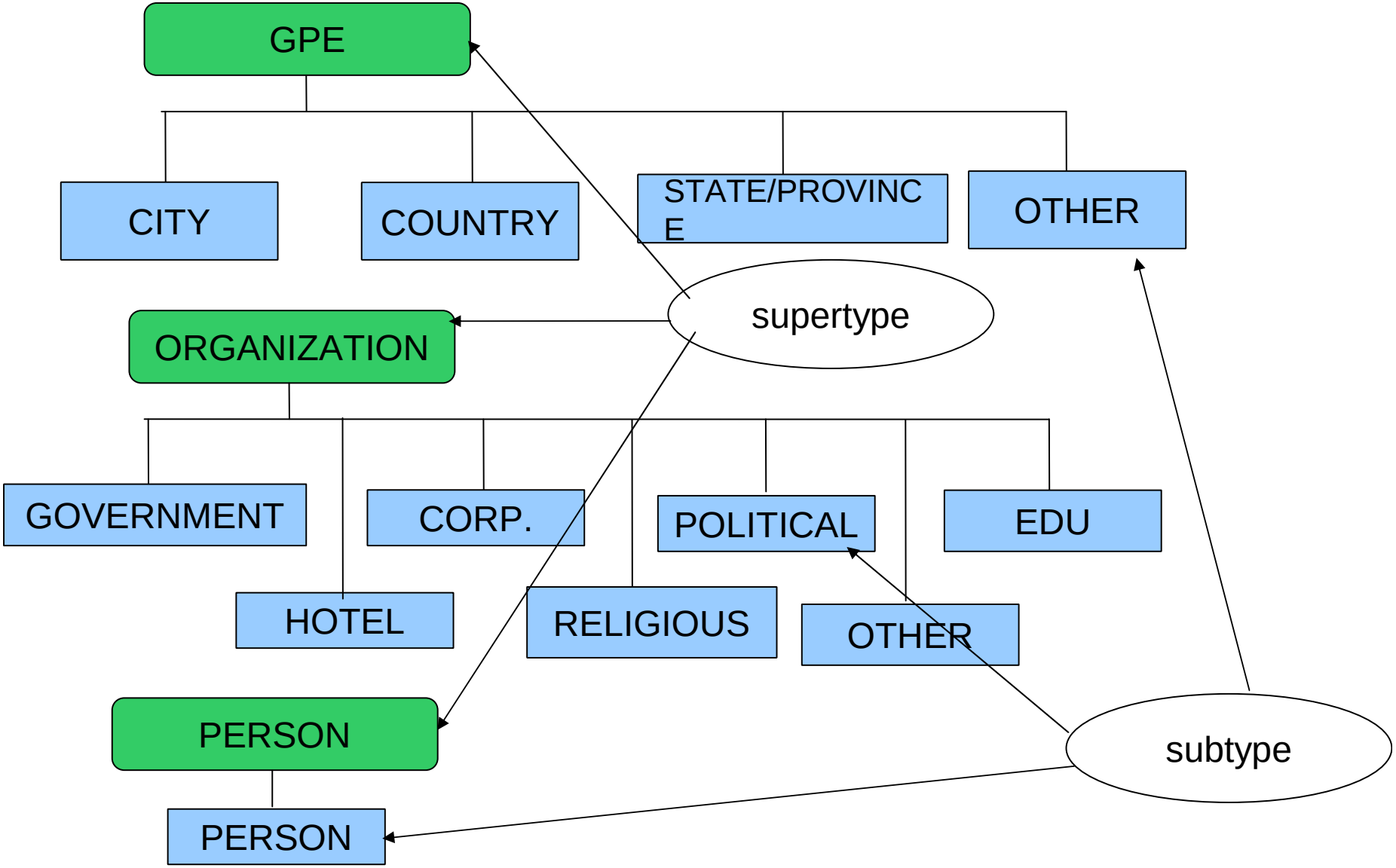
Fine-Grained Semantic Class Induction via Hierarchical and Collective Classification

Altaf Rahman and Vincent Ng
Human Language Technology Research Institute
The University of Texas at Dallas

What are semantic classes?

- PERSON, ORGANIZATION, LOCATION, FACILITY, etc ...

What are fine grained semantic classes?



Goal

- Induce semantic subtypes
  - Classify each name and nominal as one of 92 semantic subtypes predefined in the BBN Entity Type Corpus (Weischedel and Brunstein, 2005).

Subtype Examples

- FACILITY
  - Building (e.g. Twin Tower, Rockefeller Center)
  - Bridge (e.g. Golden Gate Bridge, Brooklyn Bridge)
  - Airport (e.g. DFW airport, Heathrow airport)
- ORGANIZATION
  - Government (e.g. Congress, House)
  - Corporation (e.g. Mobil Corp, IBM)
  - Political (e.g. Communist Party)
- GPE
  - Country (e.g. USA, China)
  - City (e.g. Beijing, New York City)

Supertype	Subtype	Supertype	Subtype
PERSON	person	MONEY	money
PERSON DESC	person desc	QUANTITY	1D, 2D, 3D, weight,...
NORP	nationality,religious, ...	ORDINAL	ordinal
FACILITY	building, bridge, ...	CARDINAL	cardinal
FACILITY DESC	building, bridge, ...	EVENT	war, hurricane, others
ORGANIZATION	govt, political, ...	PLANT	plant
ORG DESC	govt, political, ...	ANIMAL	animal
GPE	city, cntry, state, ...	SUBSTANCE	food, drug, chemical,..
GPE DESC	city, cntry, state, ...	DISEASE	disease
LOCATION	river, lake, sea, ...	LAW	law
PRODUCT	food, weapon, vehicle	LANGUAGE	language
PROD DESC	food, weapon, vehicle	CONTACT INFO	address, phone
DATE	date	GAME	game
TIME	time	WORK OF ART	book, play, song
PERCENT	percent		

How can we induce semantic subtypes ?

Baseline Approach

- A supervised machine learning approach
- Corpus
  - 200 WSJ articles in the BBN entity type corpus.(Weischedel and Brunstein, 2005)
- Training instance creation
  - One for each NP (name/nominal)
    - Class value is one of 92 semantic subtypes
    - Represented by 33 features

The 33 Features

7 types of features defined on each NP.

- Mention String (3)
  - house, house_2
- Verb String (3)
  - Governing verb, its sense number, semantic role…
  - go, go_1, arg1, arg2
- Semantic (3)
  - Wordnet Semantic class, synset number, NE label…
- Grammatical (2)
  - POS, ...
- Morphological (8)
  - Prefix, suffix…
- Capitalization (4)
  - All capital, Init capital, Capital Period…
- Gazetteers (8)
  - Pronouns, common words, person, vehicle, place names.

Training the baseline model

- Using Maximum Entropy
  - MaxEnt provides a probabilistic classification for each instance, which will help us to perform collective classification later on.

Improving baseline model

- Two extensions
  - Hierarchical classification
  - Collective classification

Improving baseline model

- Two extensions
  - Hierarchical classification
  - Collective classification

Hierachical Classification: Motivation

Predicting a large number of classes (92) by the baseline MaxEnt model may lead to an inaccurate estimation of the probability distribution over subtypes.

Goal :

Improve the estimation of the probability distribution over subtypes.

How ?

Hierarchical Classification

- Training
  - train a **supertype model** to classify each NP as one of 29 supertypes.
  - For each supertype train a **subtype model** to classify an NP as one of the subtypes of that particular supertype.

- Testing
  - First for each NP determine its supertype using the **supertype model**.
  - Second determine the subtype using the corresponding **subtype model**.

Training supertype and subtype models

- Feature set
  - baseline feature set
- Training instance creation
  - Supertype model
    - Same as the baseline model
  - Subtype model
    - Use only those training instances that belong to the corresponding supertype

Improving baseline model

- Two extensions
  - Hierarchical classification
  - Collective classification

Collective Classification

- Motivation
  - Problem with baseline model
    - classifies each instance independently.
    - the model cannot take into account relationships between NPs.

e.g. given string and its abbreviation should have the same semantic subtype.

  - "NYC" & "New York City"
  - But the baseline model does not enforce that they get same semantic subtype

Collective Classification

- Idea : To treat the **baseline model prediction** for each NP, which is a probability distribution as its **prior label distribution** convert it into a **posterior label distribution** by exploiting the relationship between two NPs.

- Use Factor Graphs

Factor Graph

- 2 types of node
  - Variable node. Each variable node can take one of a set of values.
  - Factor node. Each factor node is associated with a feature function that tells us the compatibility of a particular assignment of values to the nodes it connects.

Goal : Assign a value to each variable node to maximize some objective function g.

$$g(X_1, \ldots, X_n) = f_1(s_1(x_1, \ldots, x_n)) \times f_2(s_2(x_1, \ldots, x_n)) \ldots \times f_m(s_m(x_1, \ldots, x_n))$$

f_k is a feature function

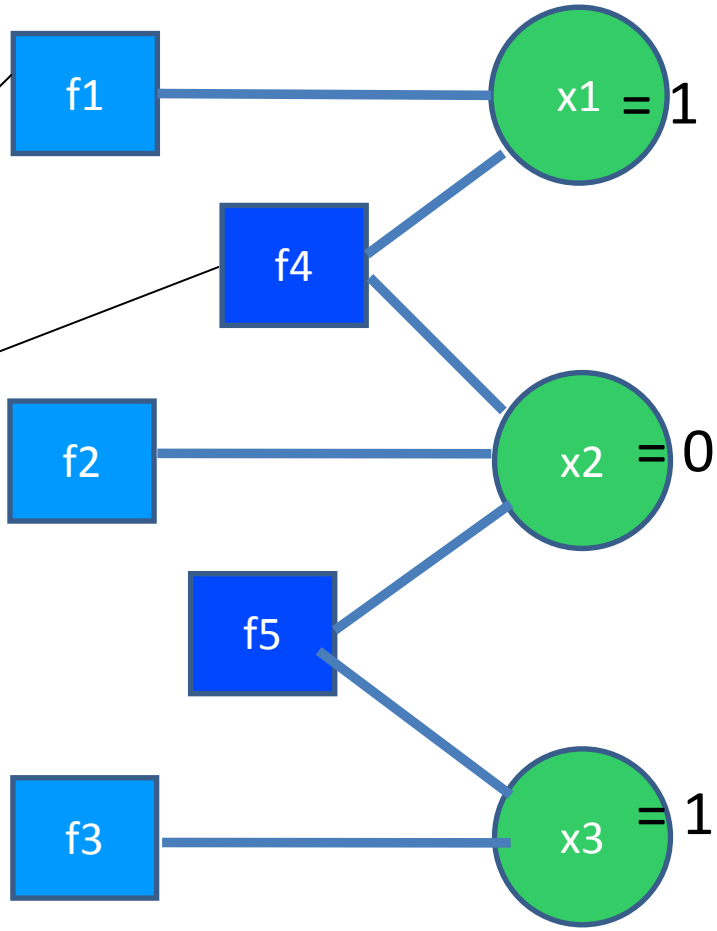
- computes the compatibility of an assignment of values to the variables in $s_k(X_1, \ldots, X_n)$

Factor Graph: An example

- Variable nodes : x1, x2, x3
- Factor nodes : f1, f2, f3, f4, f5
- Variable nodes takes the value of either 0 or 1

- Maximize an objective function g where

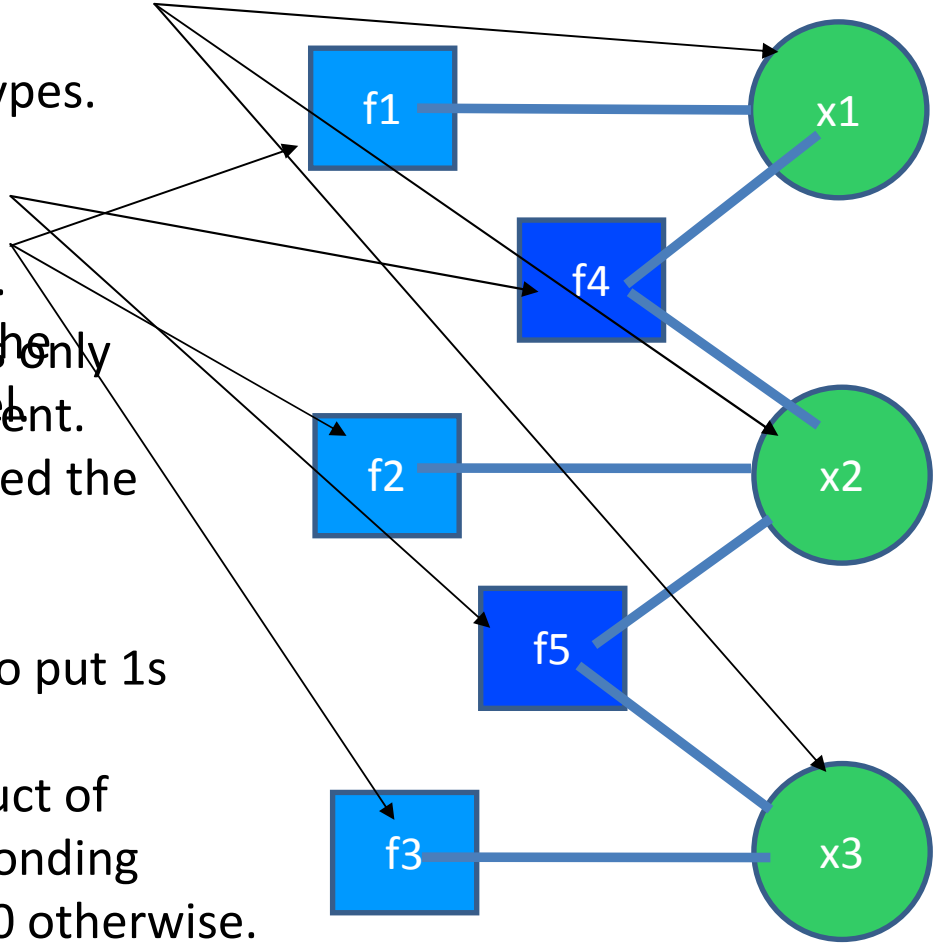
g = f1(x1) X f2(x2) X f3(x3) X f4(x1, x2) X f5(x2, x3)



f4 's table		
	x2 = 0	x2 = 1
x1 = 0	0.6	0.1
x1 = 1	0.4	0.2

Factor graph for subtype inference

- One factor graph for each test document.
- Variable node : One for each NP from that document.
  - To be assigned one of the 92 subtypes.
- Factor node :
  - Connects two variable nodes.
  - The table will have 92X92 entries.
  - Idea : Connect two variable nodes only if the corresponding NPs are coreferent.
  - Why? We want them to be assigned the same subtype.
  - How?
    - One way to fill the entries is to put 1s in diagonal and 0s other wise.
    - Better way is to put the product of the probabilities by the corresponding subtype model in diagonal and 0 otherwise.



How to determine whether two NPs are coreferent?

- Using 3 heuristics
  - They are same string (after determiners are removed)
    - "China" and "China"
  - They are aliases.
    - "New York City" and "NYC"
  - They are both proper names and at least one word in common.
    - "Delta Airlines" and "Delta"
    - "Bill Clinton" and "Hillary Clinton"

Inference

- Several methods for finding an optimal assignment of the random variables to maximize the objective function.
  - Exact inference using the sum-product algorithm (Kschischang et al., 2001).
  - Approximate inference using a belief propagation algorithm, such as loopy belief propagation.

- We choose to use loopy belief propagation as our inferencer
  - computationally more efficient than an exact inferencer.

Evaluation

- 200 Wall Street Journal Articles in the BBN Entity Type corpus
- 17,292 NPs
- 80/20 training/test split
- Baseline
- Baseline+Hierarchical
- Baseline+Collective
- Baseline+Hierarchical+Collective

Semantic Supertype	Baseline only F-measure
PERSON	90.8
PERSON DESC	89.5
SUBSTANCE	63.2
NORP	89.0
FACILITY DESC	80.0
ORGANIZATION	75.2
ORG DESC	72.8
GPE	74.7
GPE DESC	66.7
PRODUCT DESC	66.7
DATE	85.0
PERCENT	100.0
MONEY	85.3
QUANTITY	36.4
ORDINAL	100.0
CARDINAL	85.7
Overall Accuracy	**81.56**

- Supertype F-measure by micro-averaging the F-measure scores of the corresponding subtypes.
- Only 16 out of 29 types have non-zero scores are shown.
- PERSON: Good accuracy, ORG: Lower accuracy

Semantic Supertype	Baseline only F-measure	Baseline + Hierarchical (F)
PERSON	90.8	89.9
PERSON DESC	89.5	91.0
SUBSTANCE	63.2	63.6
NORP	89.0	91.3
FACILITY DESC	80.0	79.0
ORGANIZATION	75.2	75.8
ORG DESC	72.8	75.5
GPE	74.7	76.2
GPE DESC	66.7	70.0
PRODUCT DESC	66.7	66.7
DATE	85.0	85.0
PERCENT	100.0	100.0
MONEY	85.3	92.4
QUANTITY	36.4	85.0
ORDINAL	100.0	100.0
CARDINAL	85.7	87.0
Overall Accuracy	**81.56**	**82.60**

- Accuracy rises from 81.56 to 82.60.
- Error reduction 5.6%
- Statistically significant at $p = 0.04$ level.

Semantic Supertype	Baseline only F-measure	Baseline + Collective (F)
PERSON	90.8	95.9
PERSON DESC	89.5	91.1
SUBSTANCE	63.2	70.6
NORP	89.0	91.0
FACILITY DESC	80.0	73.7
ORGANIZATION	75.2	80.7
ORG DESC	72.8	74.9
GPE	74.7	74.9
GPE DESC	66.7	60.0
PRODUCT DESC	66.7	66.7
DATE	85.0	85.2
PERCENT	100.0	100.0
MONEY	85.3	85.3
QUANTITY	36.4	36.4
ORDINAL	100.0	100.0
CARDINAL	85.7	86.5
Overall Accuracy	**81.56**	**83.70**

- Accuracy rises from 81.56 to 83.70.
- Error reduction 11.6%.
- p = 0.01 level.

Semantic Supertype	Baseline only F-measure	Baseline + Both (F)
PERSON	90.8	95.8
PERSON DESC	89.5	91.0
SUBSTANCE	63.2	66.7
NORP	89.0	92.4
FACILITY DESC	80.0	79.0
ORGANIZATION	75.2	81.3
ORG DESC	72.8	75.2
GPE	74.7	81.5
GPE DESC	66.7	73.7
PRODUCT DESC	66.7	66.7
DATE	85.0	85.6
PERCENT	100.0	100.0
MONEY	85.3	93.3
QUANTITY	36.4	66.7
ORDINAL	100.0	100.0
CARDINAL	85.7	88.7
Overall Accuracy	**81.56**	**85.08**

- Accuracy from 81.56 to 85.08.
- Error reduction 19.1%,
- The difference is highly significant

($p < 0.001$).

Feature Analysis

Goal: Evaluate the contribution of the features.

- Analyzed the best performing system (baseline+both)
- Iteratively remove the features from the feature set one by one.
  - In each iteration remove the feature which showed the best accuracy without it.

Feature Analysis

Mention String	Semantic	Grammatical	Morphological	Verb String	Capitalization	Gazetteers
81.4	75.8	83.3	83.7	84.1	85.2	85.6
80.4	74.9	84.3	85.3	85.3	86.1	
80.4	78.3	83.9	86.5	86.7		
81.8	76.2	85.2	87.6			
75.4	83.4	84.6				
66.2	80.9					

- [illegible] based on the gov (87.6). verb are not useful.

Conclusion

- Two techniques for semantic subtype induction :
  - hierarchical classification
  - collective classification
- They can both significantly improve a baseline classification model.
- Applying them in combination shows even better performance.

1 Collective classification captures the relationships among subsets of instances that helped improve classification accuracy.