Glen, Glenda or Glendale: Unsupervised and Semi-supervised Learning of English Noun Gender

Shane Bergsma University of Alberta
Dekang Lin Google, Inc.
Randy Goebel University of Alberta

1. Introduction

Goal: Learn noun gender for pronoun resolution.

They asked **Glen** to take them to see **Glenda** where she lived with her **mother**. But sadly, **Glen** would not go. He did not like **Glendale**. **He** just doesn't like **Glendalians**.

⇨ **masculine feminine feminine masculine neutral plural**

⇩

He (M) : candidates: {Glendale (N), Glen (M), mother (F), Glenda (F), Glen (M)}

2. Path-based Noun Gender

Bergsma and Lin, **Learning Path-Based Pronoun Resolution**, COLING-ACL 2006:

Glen lost **his** job.
Glen says **he** intends to appeal.
They asked **Glen** for **his** help.
Glen consolidated **his** power.
Glen excused **himself.**

⇨

glen, M
glen, M
glen, M
glen, M
glen, M

⇩

Noun	M	F	N	P
glen	555	42	32	34
glenda	8	102	0	11
glendale	24	2	167	18
glendalians	0	0	0	1
glenn	3182	207	95	54
glenna	0	6	0	0

www.cs.ualberta.ca/~bergsma/Gender/

- "shane bergsma" not in data!
- No document context to disambiguate, e.g.
  "lawyer" male or female
  "Ford" a person or a car

3. Sex with Support Vector Machines

Gender Counts

Noun	M	F	N	P
glen	555	42	32	34
glenda	8	102	0	11
glendale	24	2	167	18
glendalians	0	0	0	1
glenn	3182	207	95	54
glenna	0	6	0	0

Seed Extraction

Feminine seeds
steffi graf
steinem
stella mccartney
stellar jayne
stepdaughter
stephanie
stephanie herseth
stephanie white
stepmother
stewardess

Text Corpus (3GB)

⇩

Example Extraction

Class = masculine	String = "Isinbayeva"
Contexts =	
"were shocked when **<N>** spoke out against the"	
"**<N>** 's coach and agent both"	
"**<N>** won her first major victory"	
"I would bet on Miss **<N>** being the first lady to"	

⇩

Feature Vector Creation

Context Features:

Right_1 = "won"
Right_2 = "her"
Left_1 = "Miss"
Right_* = "lady"

Type Features:

Prefix_4 = "Isin-"
Suffix_1 = "-a"
Suffix_4 = "-yeva"

⇩

Auto-Trained SVM

Semi-Supervised Training:

Auto-SVM

Gender Counts

Noun	M	F	N	P
glen	555	42	32	34
glenda	8	102	0	11
glendale	24	2	167	18
glendalians	0	0	0	1
glenn	3182	207	95	54
glenna	0	6	0	0

⇩ Features ⇩

Gold Labels ⇨ **Supervised SVM**

4. Experiments

Data:

- *Unlabeled data*: Four million noun groups / four million feature vectors
- *Labeled data:* Anaphora-annotated portion of ANC: 2.7K training set, 2.6K test set.

Results:

System	Accuracy (%)
Path-Based Gender	**91.0**
Path-Based Gender with Backoff	92.1
Auto-Trained, only *Context Features*	79.1
Auto-Trained, only *Type Features*	89.1
Auto-Trained, All Features	**92.6**
Semi-sup. Training, *Context Features*	92.4
Semi-sup. Training, *Type Features*	91.3
Semi-sup. Training, All Features	**95.5**

- Path-Based Gender systems: 63-66% on nouns with less than 10 counts
- Auto- and Semi-systems: 88% and 94% on these.
- First link ambiguous surnames (like Willey, Hill, etc.) to earlier instance (e.g. Kathleen Willey): 96.7%.