Automatic Acquisition of Gender Information for Anaphora Resolution

Shane Bergsma

May 9, 2005

Department of
Computing Science

University of Alberta

Outline of Presentation

1. Explanation of Anaphora Resolution
2. Gender/number as resolution constraints
3. Gathering gender information automatically
4. Testing learned gender on a list of <noun,gender>
5. Incorporating learned gender into a Support Vector Machine Anaphora Resolution system

Anaphora Resolution Example

"In **2004**, **Exxon Mobil** paid ***its*** **Chairman Lee Raymond** a **total** of **$38.1 million**."

- Question: "Who is the chairman of Exxon Mobil?"
- Terminology:
  - *Anaphor*: "*its*"
  - *Antecedent*: "Exxon Mobil"
- Goal: Resolve the anaphor to the correct antecedent. -- Establish *Coreference*
- Get: "Exxon Mobil's Chairman Lee Raymond"

Scope

- Third-person anaphoric pronouns, including reflexives:
  - *He, his, him, himself* (masculine)
  - *She, her, herself* (feminine)
  - *It, its, itself* (neutral)
  - *They, their, them, themselves* (plural)

Resolving Anaphora

"In **2004**, **Exxon Mobil** paid ***its*** **Chairman Lee Raymond** a **total** of **$38.1 million**."

Resolving an anaphora typically involves:

1. Parse the text to determine the noun phrases
2. Building a list of previous nouns as potential candidates
3. Filtering candidates based on gender/number agreement, grammar violations, etc.
4. Selecting most likely candidate of remaining noun based on frequency, emphasis, etc.

Gender/Number Constraints

- Provides useful constraints for resolution:

  "**John** never saw the car. ***He*** arrived late."

  – Resolve "*He*" to "John"

  "John never saw the **car**. ***It*** arrived late."

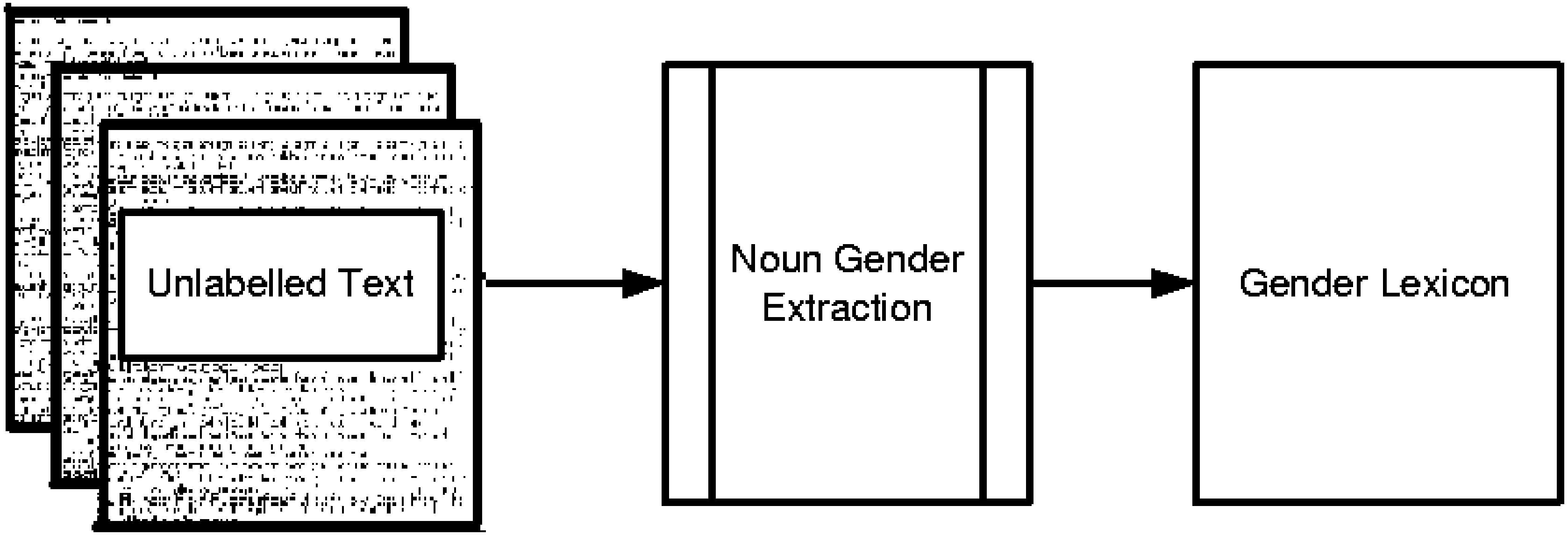
  – Resolve "*It*" to "car"

Acquiring Gender/Number Info

- Some parsers provide number information from morphology:

  "files" = "file" + "s" => "files" is plural.
- Designators: "Mr. Bean" masculine.
- Suffixes: "Chairman," "Actor," etc… not really reliable
- WordNet: Current standard – if word is subset of a person/object class, enforce restrictions

Automatic Acquistion of Gender



- Determine probability a given noun is a given gender
- Store info in gender lexicon. Use this lexicon for learning / classifying of Anaphora Resolution

The Probabilistic Approach

- Ge, Hale, and Charniak (1998) *A statistical approach to Anaphora Resolution*
  - Resolve anaphora with a simple algorithm, get gender as proportion of times noun resolved to pronoun of that gender
- Very good idea, but only 70% performance
- "Husband" found to be feminine – occurs frequently with feminine pronouns

Pattern-matching

- Example: “The ***president*** explained ***himself***.”
  - Score one for president as masculine (reflexive)
- Also pretty likely coreference:

  “The ***president*** explained ***his*** plans” (possesive)

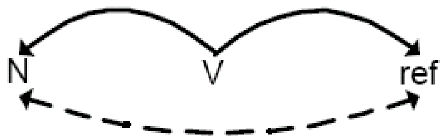
  “The ***president*** said ***he*** would explain.” (nominative)

  “***He*** is the ***president***.” (predicate)

  “Happy birthday, ***Mr***. ***President***.” (designator)
- Use a parser (Dekang Lin’s Minipar) to identify these generic situations, with any verb fillers

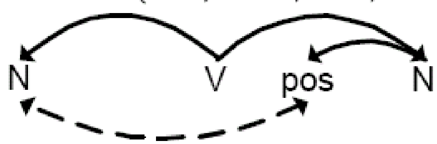
Parsed Corpus, Collect & Count:

1. Reflexives (*himself, herself, itself, themselves*):



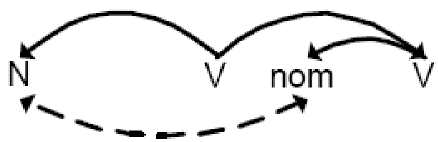
E.g. *John* explained *himself*...

2. Possessives (*his, her, its, their*):



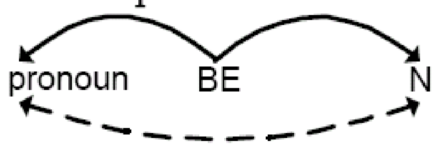
E.g. *John* bought *his* car...

3. Nominatives in *finite* sub-clauses (*he, she, it, they*):



E.g. *John* thought *he* should...

4. Predicates: pronouns are subjects and nouns are in the predicate position:



E.g. *He* is a *father*.

5. Designators: The noun is accompanied by a gendered designator:

E.g. *Mr. Johnson*.

Web-Mining Gender

- Pattern match on the *biggest* corpus (the web!)
- Use the Google API
- Count number of pages returned, e.g.:

"John * himself" "John * herself" "John * itself" "John * themselves"

- The Wildcard operator "*" substitutes for a verb
- Noisy, but effective

1. Reflexives: *himself*, *herself*, *itself*, and *themselves* in "*noun* * *reflexive*"
2. Possessives: *his*, *her*, *its*, and *their* in "*noun* * *possessive*"
3. Nominatives: *he*, *she*, *it*, and *they* in "*noun* * *nominative*"
4. Predicates: *he*, *she*, *it*, and *they* in "*nominative* is/are [a] *noun*"
5. Designators: *Mr.* and *Mrs.* in "*designator noun*"

Modelling Gender Information

- For each of the ten sources, maximum likelihood formulation:

$$P(\text{gender} = \text{masculine}) = \frac{N(\text{masculine})}{N(\text{total})}$$

- Parsed-Corpus Reflexive Count for “doctor”:

	himself	herself	itself	themselves
Count	224	126	0	14
Probability	61.5%	34.6%	0%	3.9%

Gender Usage Example

"**John** used the **computer** to access the **company's files** on ***his* purchases**."

- Resolve *his*.
- Candidates: *John, computer, company, [files]*
- WordNet contains the following senses:

"He provided *company* for her."

- (Company is a person)

"He computes faster than me. He's a good *computer*."

- (Computer is a person)

WordNet vs. Probabilistic Gender

Noun	**WordNet:** Masculine acceptable?	**Corpus Reflexives:** P(Masculine)
John	**OK**	**99.7%**
Company	**OK**	**0%** **(93% neutral)**
Computer	**OK**	**0%** **(99.2% neutral)**

Modelling Gender Information

- Have gender counts from 5 parsed corpus sources and 5 web-mined sources… how can these be combined?
- Combine the ten sources as dimensions in an feature space, learn a classifier

..... Canada..... is…

..... Canada..... she…

…n Matrix

Is "Canada" feminine? Send this vector to a classifier

			Fem	Neut	Plural
Corpus	Refl	0.00	0.00	1.00	0.00
	Pos	0.03	0.00	0.86	0.11
	Nom	0.05	0.00	0.86	0.09
	Pred	0.00	0.00	0.99	0.01
	Des	0.00	1.00	0.00	0.00
Web	Refl	0.12	0.08	0.59	0.11
	Pos	0.11	0.05	0.66	0.18
	Nom	0.18	0.05	0.61	0.16
	Pred	0.05	0.02	0.28	0.65
	Des	0.70	0.30	0.00	0.00

Further Considerations

What about sparse data?

- Use add-one smoothing to deal with low counts, that is, initially assume each gender was seen once

What about confidence in counts?

- Quantify that 25% for a noun never seen is less confident than 25% for a noun seen six thousand times – i.e. want variance measure
- Add-one smoothing + variance measure = Beta distribution

Outline of Presentation

- ✓ Explanation of Anaphora Resolution
- ✓ Gender/number as resolution constraints
- ✓ Gathering gender information automatically

4. Testing learned gender on a list of <noun,gender>
5. Incorporating learned gender into a Support Vector Machine Anaphora Resolution system

Data Sets

- Labelled 2779 pronouns from news articles in the American National Corpus (data set is available):

  “the dog likes <coref ante=dog>its</coref> toy”
- Divided into Training Set / Test Set
- Also use these tags to extract within-context gender of nouns:

  “<coref ante=dog>its</coref>” => <dog, neutral>
- Build a gendered-noun list

Task#1: Guess Noun Gender

- For each word in gendered noun list, use probabilities to guess gender
- 5 corpus and 5 web-mined Beta-distribution means (i.e. probability guesses) and variances provide a 20-element feature vector.
- Support Vector Machines (SVM^{light}, linear kernel) learn separate classifiers for masc/fem/neut/plural
- Learn on Training Set nouns, test on Test set

One List, Four Gender Classifiers

Masc SVM

Fem SVM

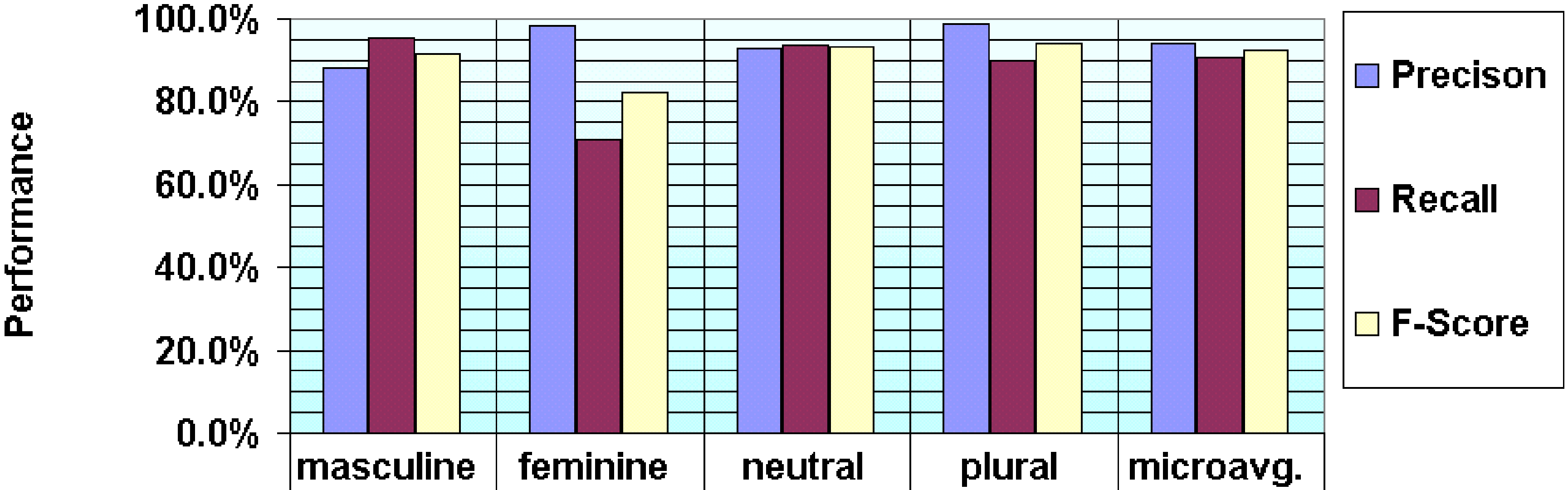
Fem

Fem

Fem SVM

Word	Masc	Fem	Neut	Plural
buffett	T	F	F	F
stock	F	F	T	F
wife	F	T	F	F
magazines	F	F	F	T
tripp	F	T	F	F

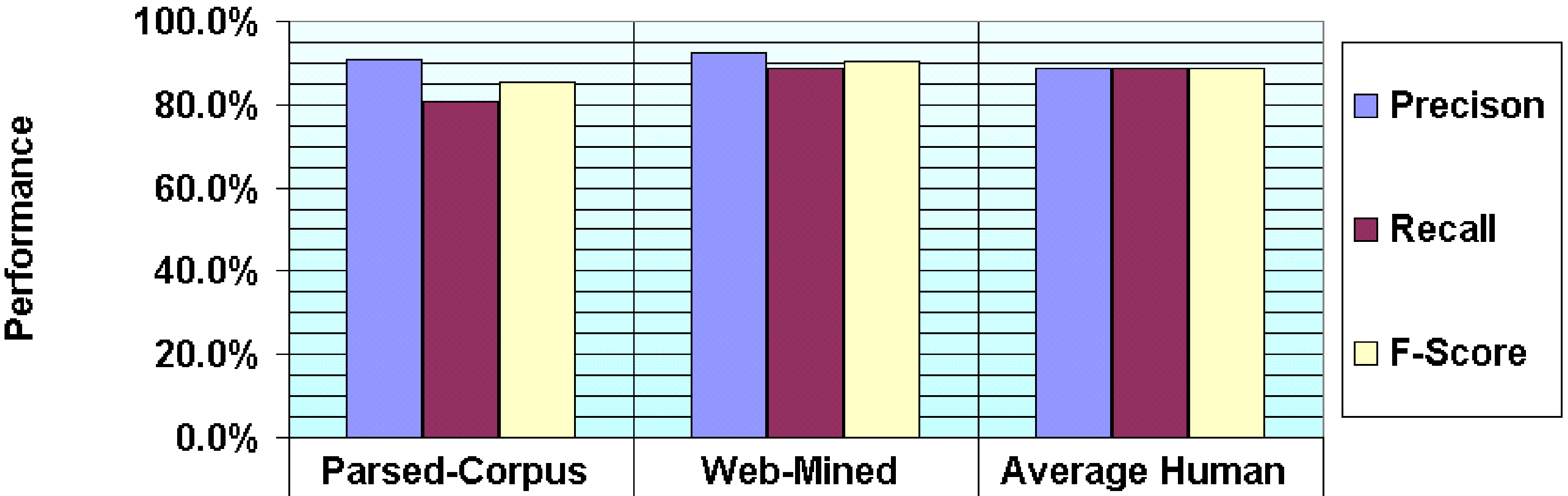
Overall Classification Performance



	masculine	feminine	neutral	plural	microavg.
Precison	88.2%	98.2%	93.0%	98.6%	93.9%
Recall	95.2%	70.9%	93.7%	89.8%	90.6%
F-Score	91.6%	82.4%	93.3%	94.0%	92.2%

Pronoun category

Special Classification Performance



	Parsed-Corpus	Web-Mined	Average Human
Precison	90.9%	92.4%	88.8%
Recall	80.6%	88.6%	88.8%
F-Score	85.4%	90.4%	88.8%

Pronoun category

Test#2 Full Anaphora Resolution

"**Bob** had **wine** with ***his*** **supper**."

Five systems for comparison:

1. Baseline – Choose previous noun
2. Baseline with hard gender constraints
3. Baseline with hard + probabilistic gender constraints
4. Full SVM system with hard gender constraints
5. Full SVM system with hard gender constraints + probabilistic gender constraints

Machine Learning

- ML approach to Anaphora Resolution:
  - Each instance is a candidate noun/anaphor pair, and classifier decides if corefent.
  - Apply classifier backward incrementally until antecedent is accepted

Bob had wine with <coref ante=Bob>his</coref> supper. <wine,his>:(−ve)

Bob had wine with <coref ante=Bob>his</coref> supper. <Bob,his>:(+ve)

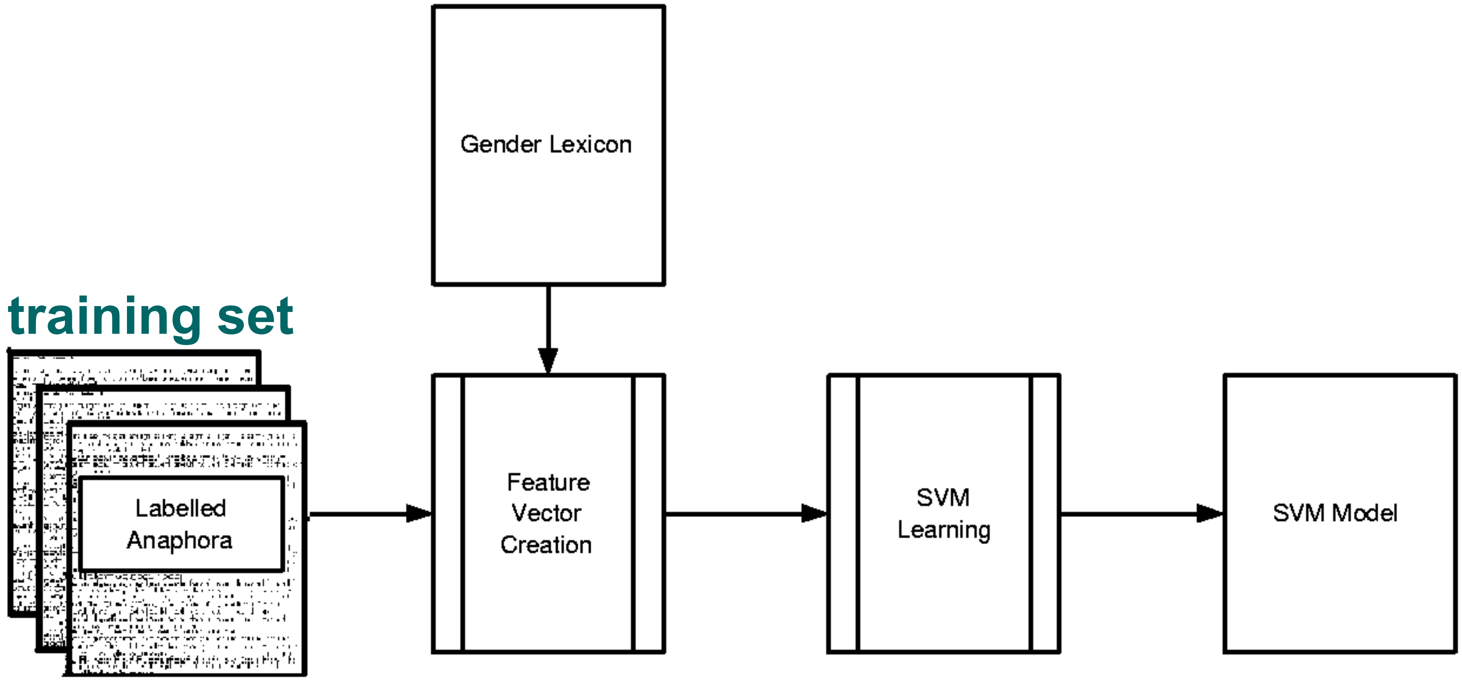
Feature vectors

- A number of features made available to SVM
- Features for the Pronoun, the Antecedent, the pair, and Gender

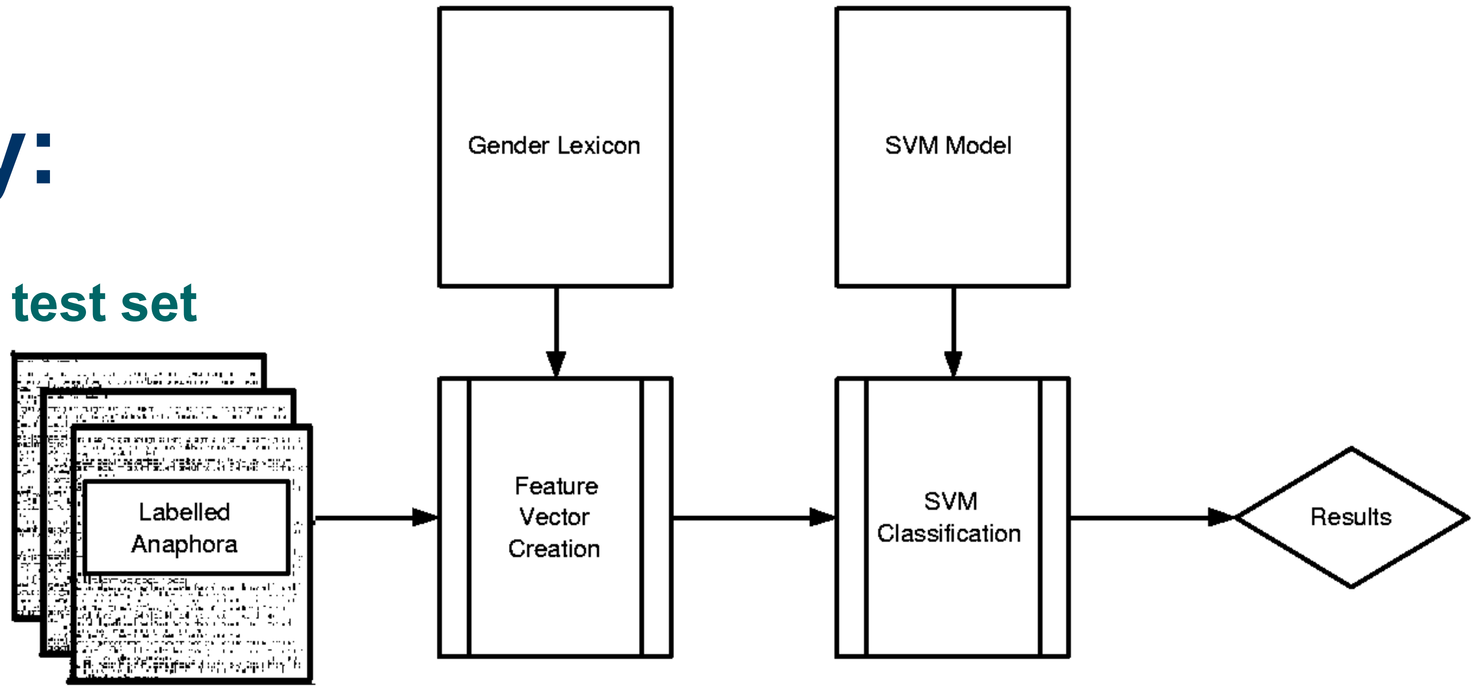
Table 4. Features for Pronoun Resolution

Type	Feature	Description
Pronoun Features	Masculine	1: pronoun masculine; else 0
	Feminine	1: pronoun feminine; else 0
	Neutral	1: pronoun neutral; else 0
	Plural	1: pronoun plural; else 0
Antecedent Features	Antecedent Frequency	Number of Occurrences / 10.0
	Subject	1: subject of clause; else 0
	Object	1: object of clause; else 0
	Predicate	1: predicate of clause; else 0
	Pronominal	1: pronoun; else 0
	Prepositional	1: prepositional complement; else 0
	Head-Word Emphasis	1: parent not noun; else 0
	Conjunction	1: *not* part of conjunction; else 0
	Prenominal modifier	1: noun is a prenominal modifier; else 0
	Org	1: an organization; else 0
	Person	1: a person; else 0
	Time	1: has time units; else 0
	Date	1: a date; else 0
	Money	1: a monetary denomination; else 0
	Price	1: a price; else 0
	Amount	1: ante has measurement units; else 0
	Number	1: number; else 0
	Definite	1: has definite article; else 0
	His/Her	1: ante first word of his/her pattern; else 0
	He/His	1: ante first word of he/his pattern; else 0
Gender Features	Std. Gender Match	1: gender known and matches; else 0
	Std. Gender Mismatch	0 if gender known and mismatches; else 1
	Pronoun Mismatch	0 if both pronouns and mismatch; else 1
	Web/Corpus Genders	mean/std. dev. of *Beta* distributions (20X)
Pronoun-Antecedent Features	Binding Theory	1: satisfies Principles B,C; else 0
	Reflexive Subj. Match	1: ante subj. of reflexive pron's GC; else 0
	Same Sentence	1: ante/pron in same sentence; else 0
	Intra-Sentence Diff.	Within-sentence difference/50.0
	In Previous Sentence	1: ante in previous sentence; else 0
	Inter-Sentence Diff.	Sentence distance/50.0
	Prepositional Parallel	1: ante/pron objs. of same preposition; else 0
	Relation-Match	1: ante/pron have same gramm. rel.; else 0
	Parent Relation Match	1: parents have same gramm. rel.; else 0
	Parent Cat. Match	1: parents have same gramm. category; else 0
	Parent Word Match	1: parents same word; else 0
	Quotation Situation	1: ante/pron both in/out of quotes; else 0
	Singular Match	1: both singular; else 0
	Plural Match	1: both plural; else 0
	MI Value	Mutual Information between ante and pron
	MI Available	1: MI value available; else 0

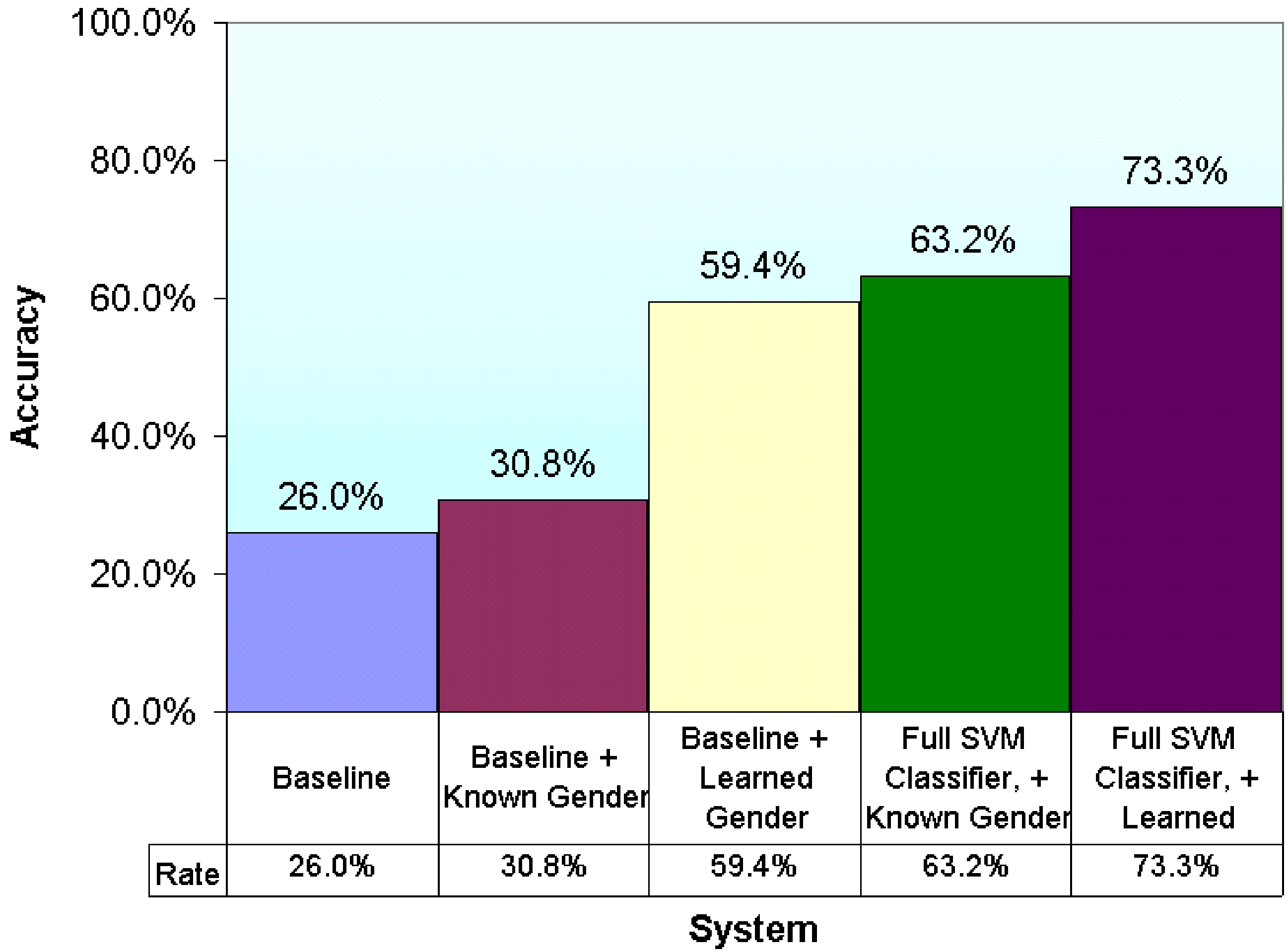
Learn:



Classify:



Pronoun Resolution Performance



Conclusions

- 73.3% competitive with other systems with automatic noun-identification, parsing (Kennedy & Boguraev, 75%, Mitkov, 62%)
- Gender-guessing outperforms humans
- Parsed corpus and web features work together
- Learned gender shown to result in *significant* performance improvements over standard gender approach

Future Work

- Better parsing on more text, larger world wide web – all will automatically help our approach
- Recent developments:
  - using EM to learn gender and make resolutions in a large text completely unsupervised
  - Mining other information from text – including likelihood of coreference across syntactic relations

Gratitude

- Thank you very much for your time and attention
- Thank you to my supervisor, Dr. Dekang Lin
- Thanks to NSERC and iCORE for funding
- Questions

NSERC
CRSNG

iCORE