Entities, Coreference, Discourse, etc.

Hal Daumé III

School of Computing
University of Utah

me@hal3.name

About Summarization Research

- Sentence Extraction
  - Given a document, identify and extract *important* sentences
  - Special models to handle redundancy
  - Typically trained on manually annotated or automatically annotated extracts

There are a wealth of document/abstract pairs that statistical summarization systems could leverage to learn how to create novel abstracts. Detailed studies of such pairs~\cite{jing:cl} show that human abstractors perform a range of very sophisticated operations when summarizing texts, which include reordering, fusion, and paraphrasing. Unfortunately, existing document/abstract alignment models are not powerful enough to capture these operations. To get around directly tackling this problem, researchers in text summarization have employed one of several techniques.

Some researchers~\cite{bankoetal00} have developed simple statistical models for aligning documents and headlines. These models, which implement IBM Model 1~\cite{brownetal93}, treat documents and headlines as simple bags of words and learn probabilistic word-based mappings between the words in the documents and the words in the headlines. As our results show, these models are too weak for capturing the operations that are employed by humans in summarizing texts beyond the headline level.

Other researchers have developed models that make unreasonable assumptions about the data, which lead to the utilization of a very small percent of available data. For instance, the document and sentence compression models of Daum\'e III, Knight, and Marcu~\cite{knight-marcu02,daume-marcu02} assume that sentences/documents can be summarized only through deletion of contiguous text segments. Knight and Marcu found that from a corpus of $39,060$ abstract sentences, only 1067 sentence extracts existed: a recall of only 2.7%.

About Actual Summaries

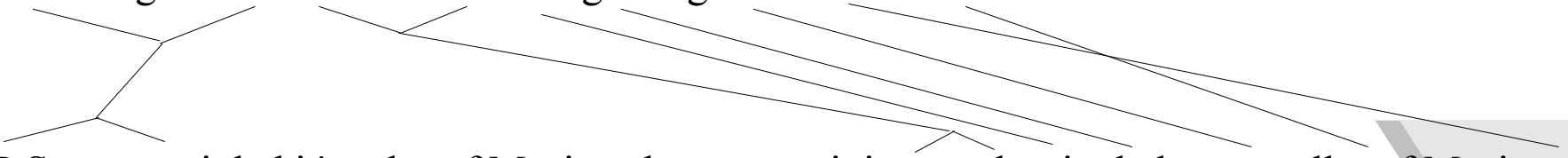
- Are *not* extracts
  - The sentence is not an appropriate level of granularity
- Are *not* compressions
  - Involve lots of rewriting and reordering
- Are *not* bags of words
  - Are fluent, grammatical, etc.
- **So why do we focus on these unrealistic problems?**

DATA!

Learning Transformations

- Document/Abstract pairs:

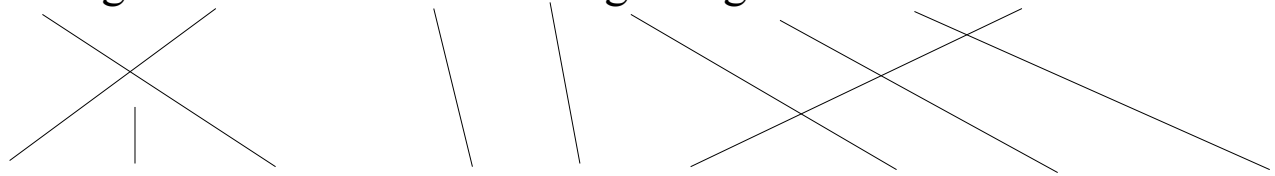
Connecting Point has become the single largest Mac retailer.



CP Systems tripled it's sales of Macintosh systems; it is now the single largest seller of Macintosh.

- English/French pairs:

Connecting Point has become the single largest Mac retailer.



L' Pointe de Connecting bécomé l' retailerese oné-most largezze Macintosh.

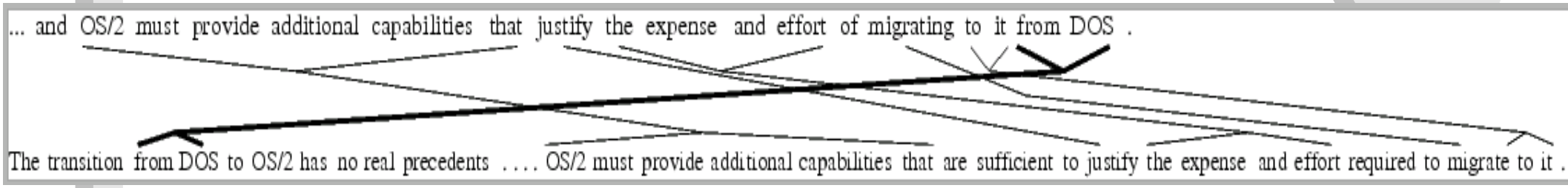
- How does MT solve this problem?

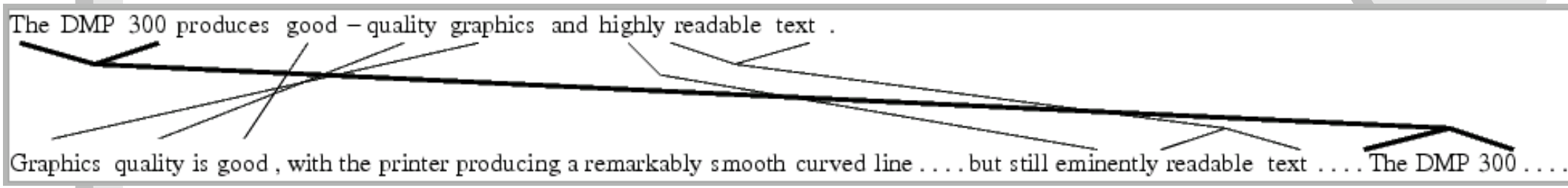
Alignments!!!

Results

System	Precision	Recall	F-Score
Human 1	0.727	0.746	0.736
Human 2	0.680	0.695	0.687
GIZA-HMM	0.120	**0.260**	0.164
GIZA-Model 4	0.117	**0.260**	0.161
GIZA-HMM (flipped)	**0.295**	0.250	**0.271**
GIZA-Model 4 (flipped)	0.280	0.247	0.262
Decomposition	**0.349**	**0.379**	**0.363**
PBHMM	**0.456**	**0.686**	**0.548**

Sources of Error





Entity Detection and Tracking

Terri Schiavo

Name

Terri Schiavo: 15 years of questions and uncertainty

Pronoun

In February 1990, cardiac arrest deprived Terri Schiavo of oxygen to her brain for five minutes - five minutes that have led to years of emotional distress and legal battles. There was initial hope for recovery, but there came a point at which the views of Terri's future diverged. In 1998, her husband, Michael Schiavo, filed the first petition to remove Terri's feeding tube and allow her to die. Since then, Terri's future has been fought over in the courts until a judge once again ordered the feeding tube removed Oct. 15, 2003. Legal avenues exhausted, Bob and Mary Schindler, Terri's parents, turned to the Florida Gov. Jeb Bush and then to the Florida legislature, which passed a bill allowing the governor to order Terri Schiavo's feeding tube be reinserted.

Premodifier

Nominal

Entity Detection and Tracking

- Official formulation:
  - Identify all *entities* appearing in a document and the textual spans (*mentions*) that refer to these entities
- Typical interpretation:
  - Identify all *mentions* appearing in a document and discern which mentions refer to the same *entity*
- Identifying mentions also involves mention types:
  - Name (NAM), Nominal (NOM), Pronoun (PRO), Premodifier (PRE)
- Identifying entities also involves entity types:
  - Person, Organization (+5 subtypes), GPE (+10), Location (+6), Facility (+8), Vehicle (+5), Weapon (+9)

Mention Detection

- Use BIO-encoding to obtain sequence labeling problem:

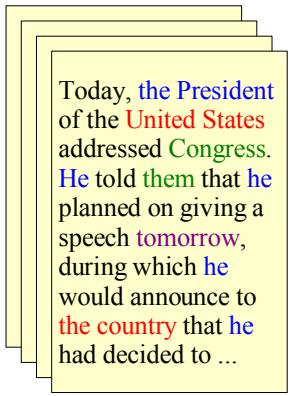
ordered the feeding tube removed Oct. 15, 2003. Legal avenues exhausted, Bob and Mary Schindler, Terri's parents, turned to the Florida Gov. Jeb Bush and then to the Florida legislature, which passed a bill allowing the governor to order Terri Schiavo's feeding tube be reinserted.

...	Bob	and	Mary	Schindler	,	Terri	's	parents	,
	B-Per	I-Per	I-Per	I-Per	O	B-Per	O	B-Per	O

turned	to	the	Florida	Gov.	Jeb	Bush	and	then	...
O	O	O	B-GPE	B-Per	B-Per	I-Per	O	O	

- Now, *if we make a Markov assumption on the labels*, we can apply any standard sequence labeling model:
  - HMM, MEMM, CRF, M3N, ...

Coreference Resolution

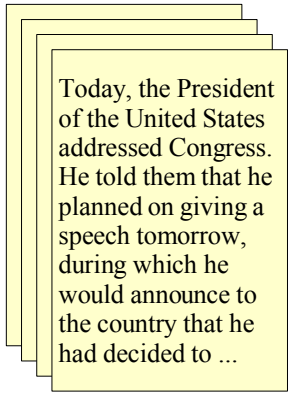


Original Training Data

-	the President	United States
-	the President	Congress
+	the President	he
-	the President	them
	...	
-	United States	Congress
-	United States	he
+	United States	the country
	...	
-	the country	he
	...	

Binary Training Pairs

Your Favorite ML Algorithm



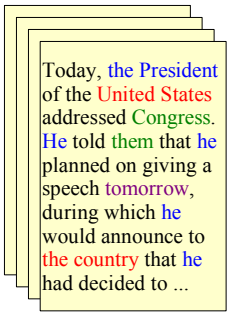
Test Data

?	the President	United States
?	the President	Congress
?	the President	he
?	the President	them
	...	
?	United States	Congress
?	United States	he
?	United States	the country
	...	
?	the country	he
	...	

Binary Test Pairs

f

-	**the President**	**United States**
-	the President	Congress
+	**the President**	**he**
-	the President	them
	...	
-	United States	Congress
+	**United States**	**he**
+	United States	the country
	...	
-	the country	he
	...	



Your Favorite Clustering Algorithm

Coreference Resolution

- Choosing a classifier
  - How should we tune it?
  - Can we train on all pairs?
- Choosing training instances
  - Use all pairs? Most recent negatives only? Samples?
  - What about the i.i.d. assumption?
- Choosing a clustering algorithm
  - How does this choice interact with the classifier?
  - How can we tune parameters?

Two Successful Approaches

- Classifier:
  - Maxent multilabel classifier
- Instances:
  - All pairs
- Clustering:
  - Use a max-link beam search (the *Bell-tree* algorithm)

[Florian et al., 2004]

- Classifier:
  - Perceptron-trained CRF
- Instances:
  - All pairs
- Clustering:
  - Use generic graph partitioning algorithms

[McCallum & Wellner, 2004]

Summary of Previous Approaches

- Mention Detection:
  - Tractable under Markov assumption
  - Inference requires evaluation of forward/backward (sum-product) algorithm for likelihood or margin-based training
    - For perceptron training, requires evaluation of Viterbi algorithm
  - Prediction requires evaluation of Viterbi (max-product) algorithm

- Coreference Resolution:
  - Full inference is never tractable
  - Only the McCallum & Wellner model solves the problem directly
    - But has to resort to very simple Perceptron-style updates
  - For the most part, non-integrated classification + clustering
  - Features are usually simple pairwise-comparisons

Features

Lexical: **unigrams** (words); the **bigrams**; the two character **prefixes** and **suffixes**; the word **stem**; the **case** of the word, computed by regular expressions.

Syntactic: unigrams and bigrams of **part of speech** as well as **shallow-parse** features.

Semantic: two most common **synsets**; all **hypernyms**; for coreference, **distance** in the WordNet graph between pairs of head words whether one is a **part of** the other; synset and hypernym information of the **preceding** and **following verbs**

Lists: about 40 lists of common places, organization, names, etc.

Class: word clusters

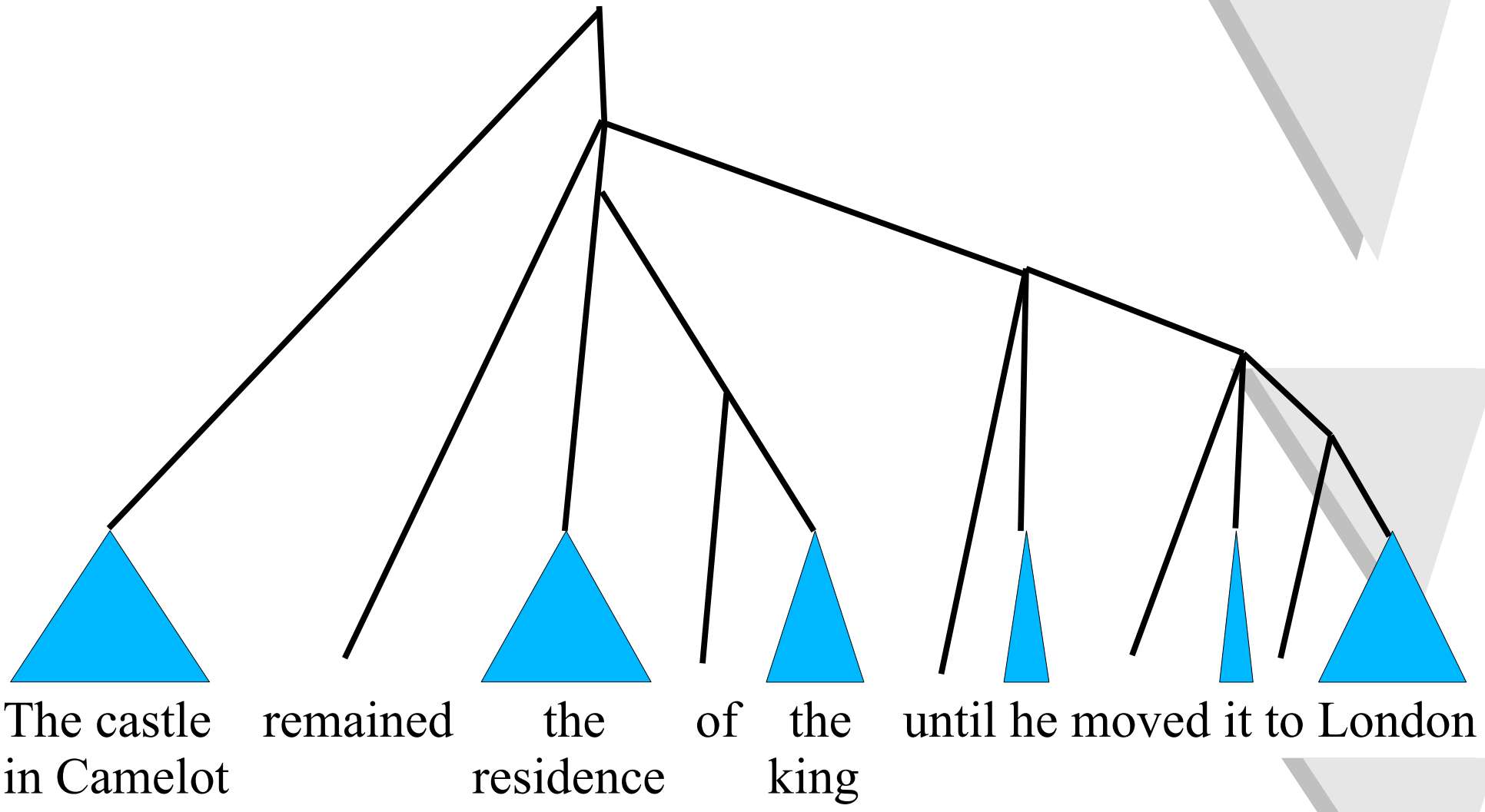
Inference: models to predict **number** and **gender**; output of MEMMs trained off of the **MUC6**, the **MUC7** and **ACE** data.

String: string **match**; **substring** match; string **overlap**; **pronoun** match; and normalized **edit distance**; string **nationality** match; linguistically-motivated string edit distance; **Jaro** distance; **acronym** match.

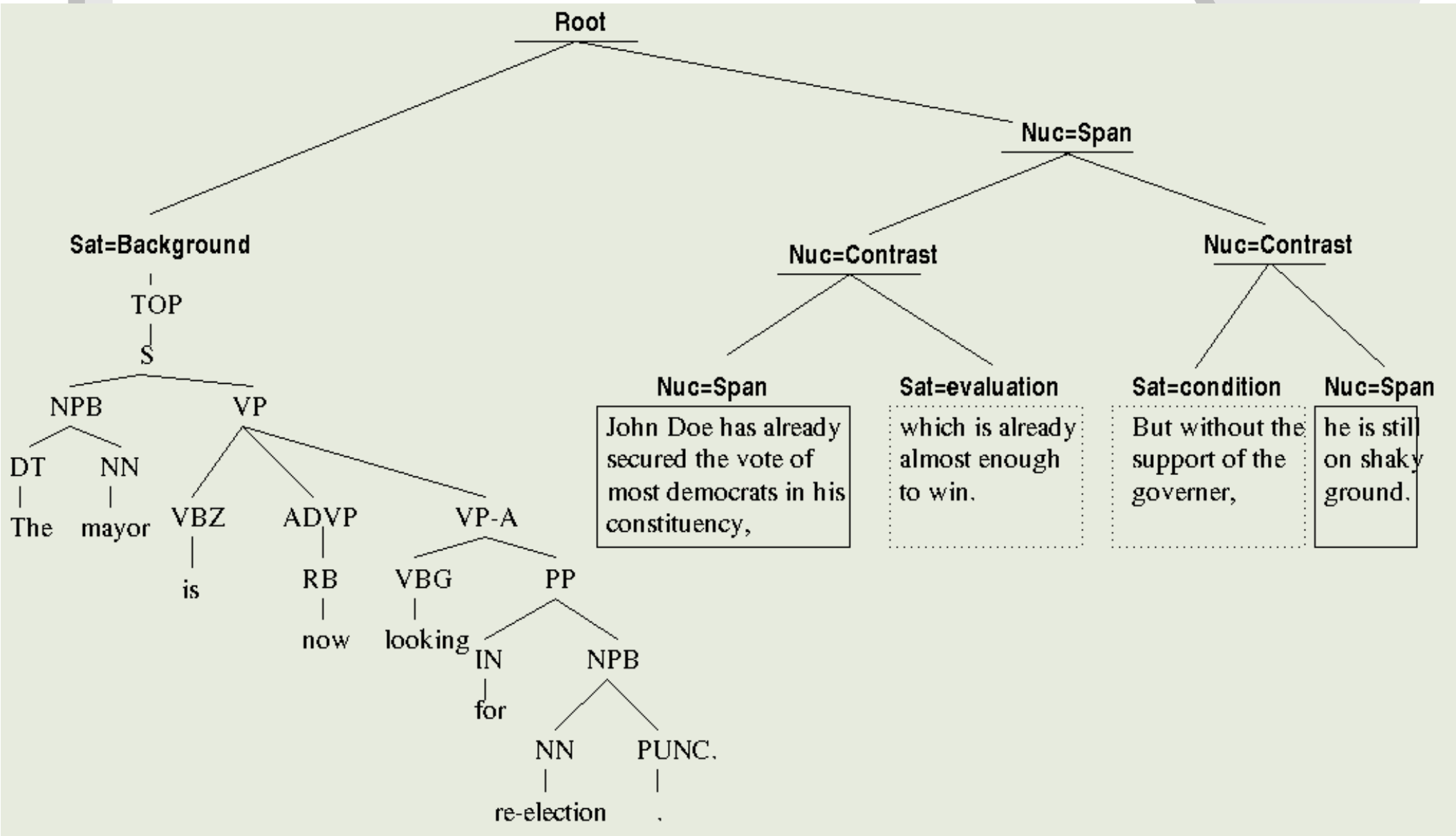
Performance

Name to Name	mid to high 90s
Name to Pronoun	mid 80s
Name to Nominal	40s-50s

Hobbs' Distance



Discourse "Hobbs' Distance"



Does Discourse Distance hold?

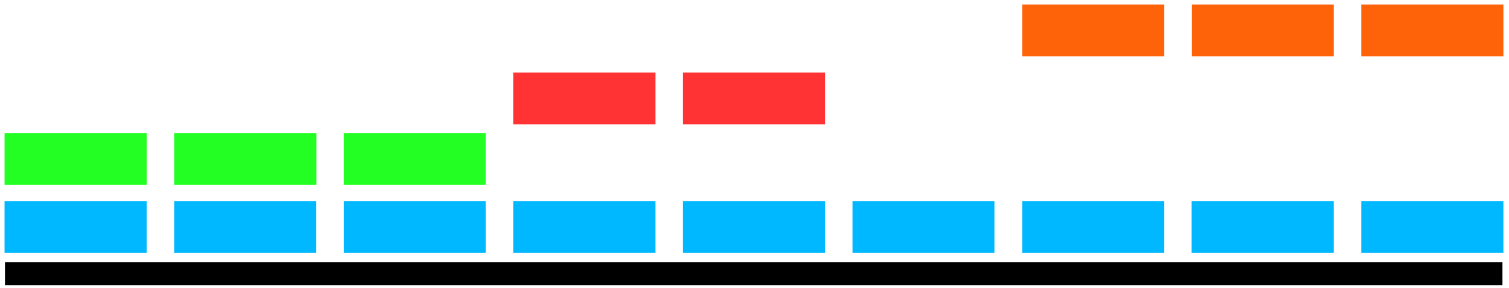
Holds perfectly in ~ 55%

60% of remaining are ATTRIBUTION

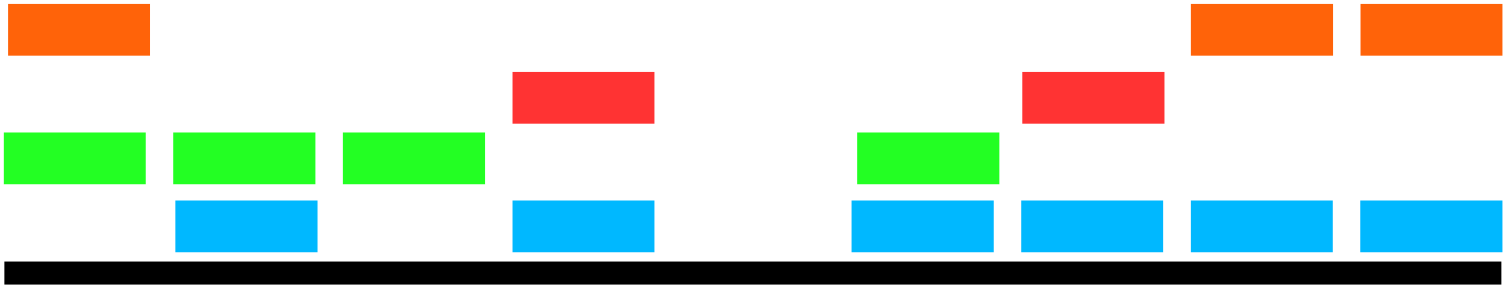
85% of remaining binuclear

Distribution of Entities

Expectation



Truth



Name/Instance Data

12416	Palestinian leader	Yasser Arafat
5772	Bosnian Serb leader	Radovan Karadzic
3839	White House spokesman	Mike McCurry
3660	Foreign editor	Rick Christie
3654	News editor	Art Dalglish
3228	State Department spokesman	Nicholas Burns
3089	White House spokesman	Marlin Fitzwater
2528	PLO leader	Yasser Arafat
2157	first lady	Hillary Rodham Clinton
2069	spokesman	Alexander Ivanko
1677	Soviet leader	Mikhail Gorbachev
1646	envoy	Richard Holbrooke
1585	Libyan leader	Moammar Gadhafi
1196	envoy	Dennis Ross
1195	Communist leader	Gennady Zyuganov
1152	White House spokesman	Joe Lockhart
1109	Turkish Cypriot leader	Rauf Denktash
1057	White House press secretary	Mike McCurry

Using Name/Instance Data

ordered the feeding tube removed Oct. 15, 2003. Legal avenues exhausted, Bob and Mary Schindler, Terri's parents, turned to the Florida Gov. Jeb Bush and then to the Florida legislature, which passed a bill allowing the governor to order Terri Schiavo's feeding tube be reinserted.

Gazetteers

- **census** data and **baby name** books
- standard **gazetteers**
  (countries, cities, islands, ports, provinces and states)
- airport locations
- company names (**NASDAQ** and **NYSE**)
- s**emantically plural** words
- list of persons, organizations and locations
  that were identified by **IdentiFinder**.

Using Gazetteers for Coref

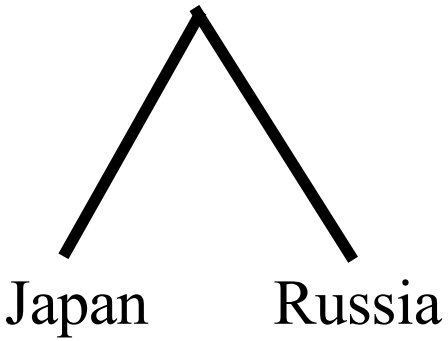
ordered the feeding tube removed Oct. 15, 2003. Legal avenues exhausted, Bob and Mary Schindler, Terri's parents, turned to the Florida Gov. Jeb Bush and then to the Florida legislature, which passed a bill allowing the governor to order Terri Schiavo's feeding tube be reinserted.

list_of(anaphor) + lexeme_of(antecedent)

list_of(anaphor) + list_of(antecedent) + same_word?

WordNet

Distance in graph?



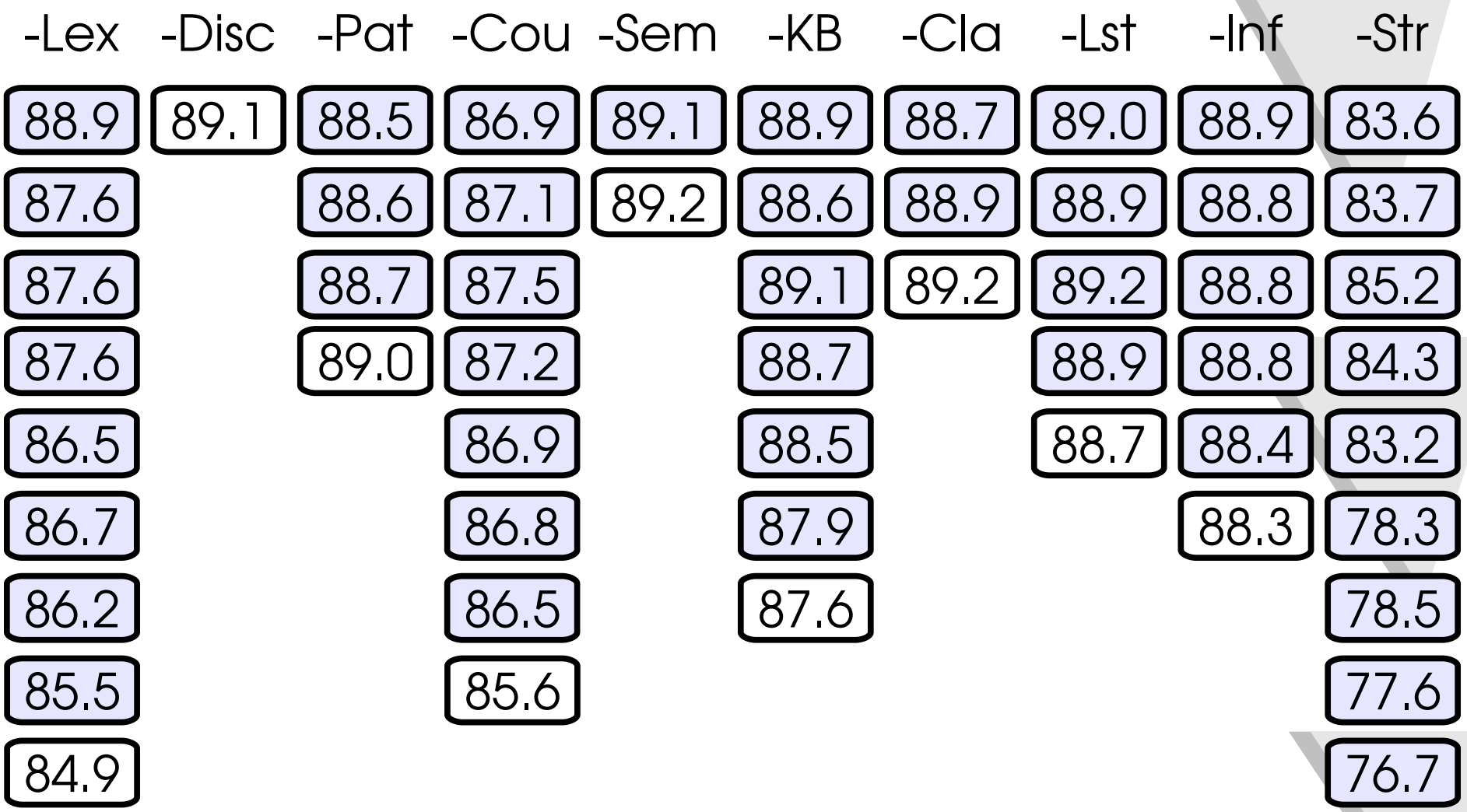
Hyponym/Hypernym?

Nearest-verb hypernyms

Count-based Features

- total **# of entities** detected
- total **# of mentions**
- Ratios:
  - **entity:mention** ratio
  - **entity:word** ratio
  - **mention:word** ratio
  - # of **mentions in the current chain to the total # of mentions**
- **size** of the hypothesized chain
- # of **intervening** mentions
- # of **intervening** mentions of the same type; # of intervening **sentence** breaks
- decayed density

Feature Contributions



Str > Lex > Cou > KB > Inf > Lst > Pat > Cla > Sem > Disc

Linkage Types

- When hypothesizing merging a mention into a chain, to which chain element to we 'attach?'

Terri Schiavo

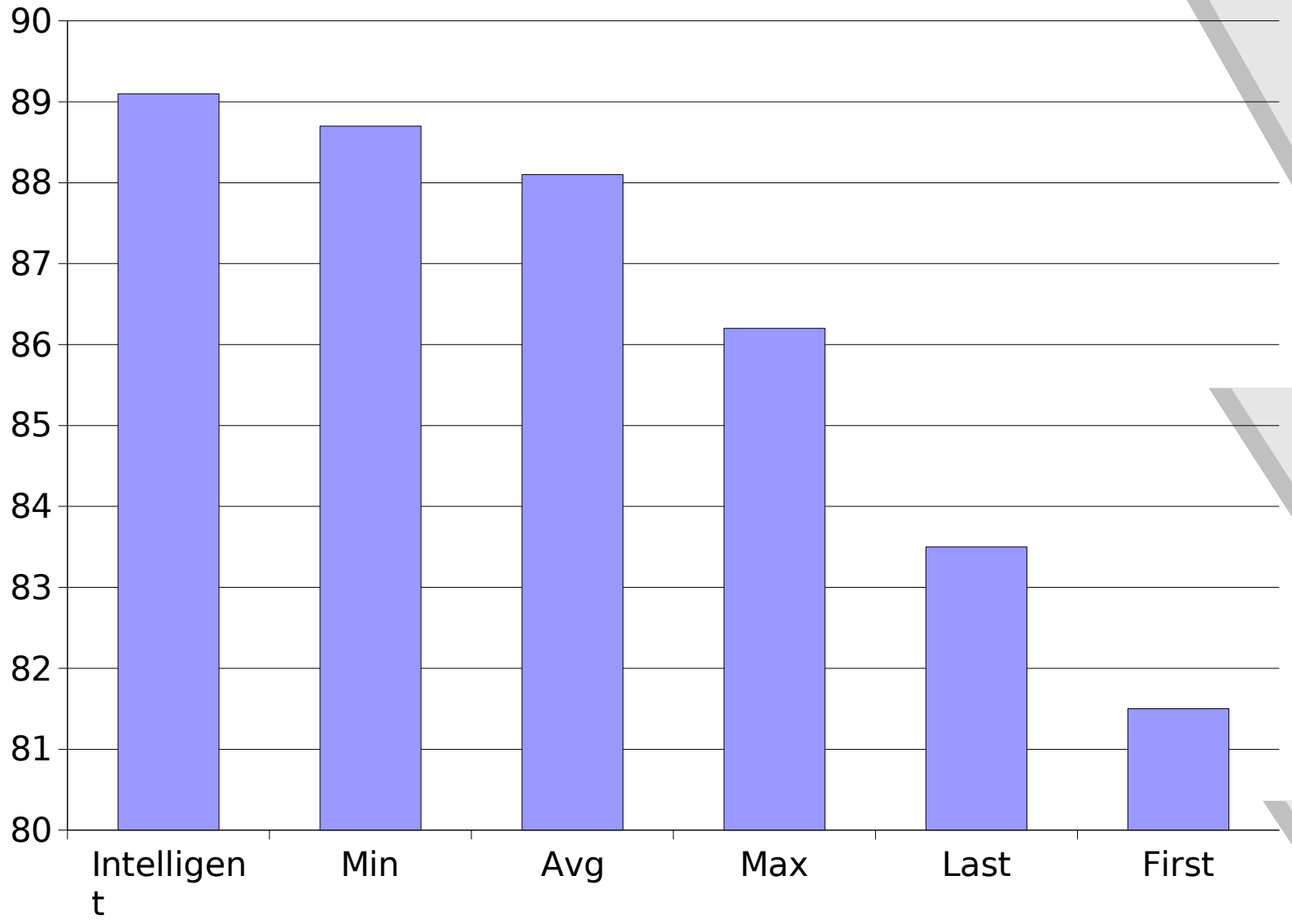
Terri Schiavo: 15 years of questions and uncertainty

In February 1990, cardiac arrest deprived Terri Schiavo of oxygen to her brain for five minutes - five minutes that have led to years of emotional distress and legal battles. There was initial hope for recovery, but there came a point at which the views of Terri's future diverged. In 1998, her husband, Michael Schiavo, filed the first petition to remove Terri's feeding tube and allow her to die.

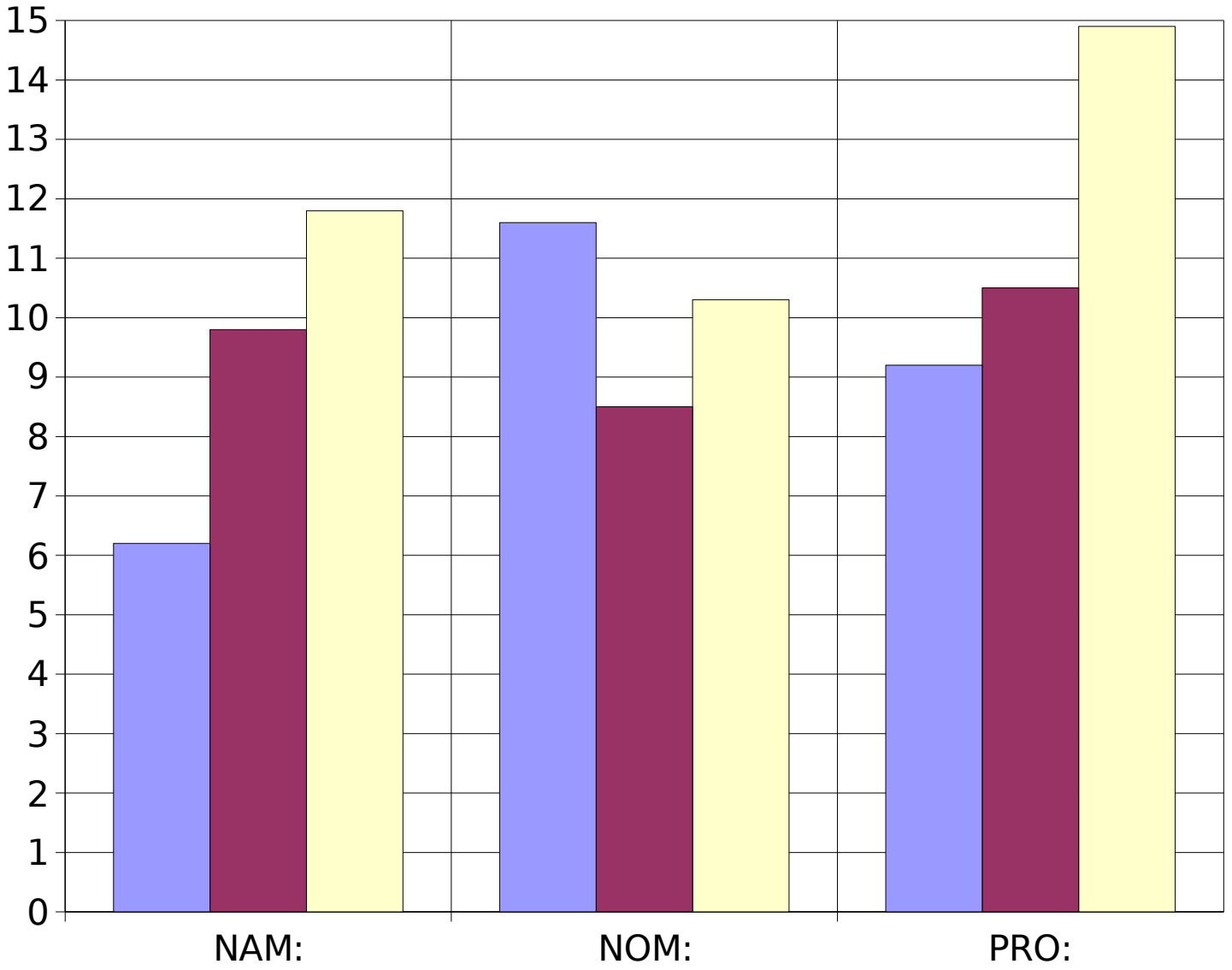
Sum/Average link (default)
Min link
Max link (commonly used)
Last link (commonly used)
First link

Intelligent link:
NAM: first(NAM) + last(NOM) + max
NOM: max(NOM) + last(NAM) + max
PRO: avg(PRO+NAM) + max

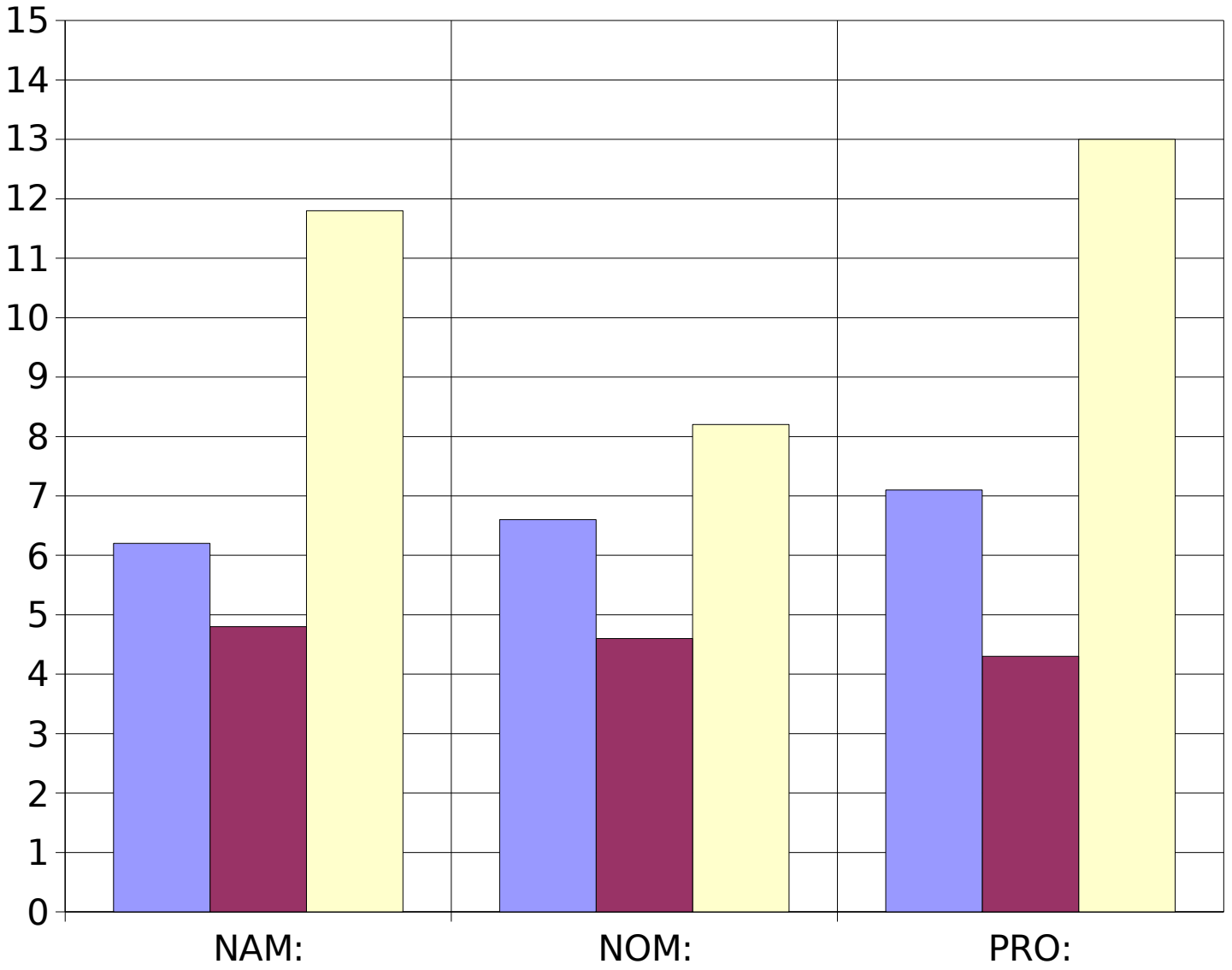
Linkage Types



Errors Pre-Engineering



Errors Post-Engineering



Discussion

Learning makes a difference (LaSO → Searn)

Feature engineering makes a bigger one

Many features not obvious
⇉ LOOK at outputs!

Be clever!