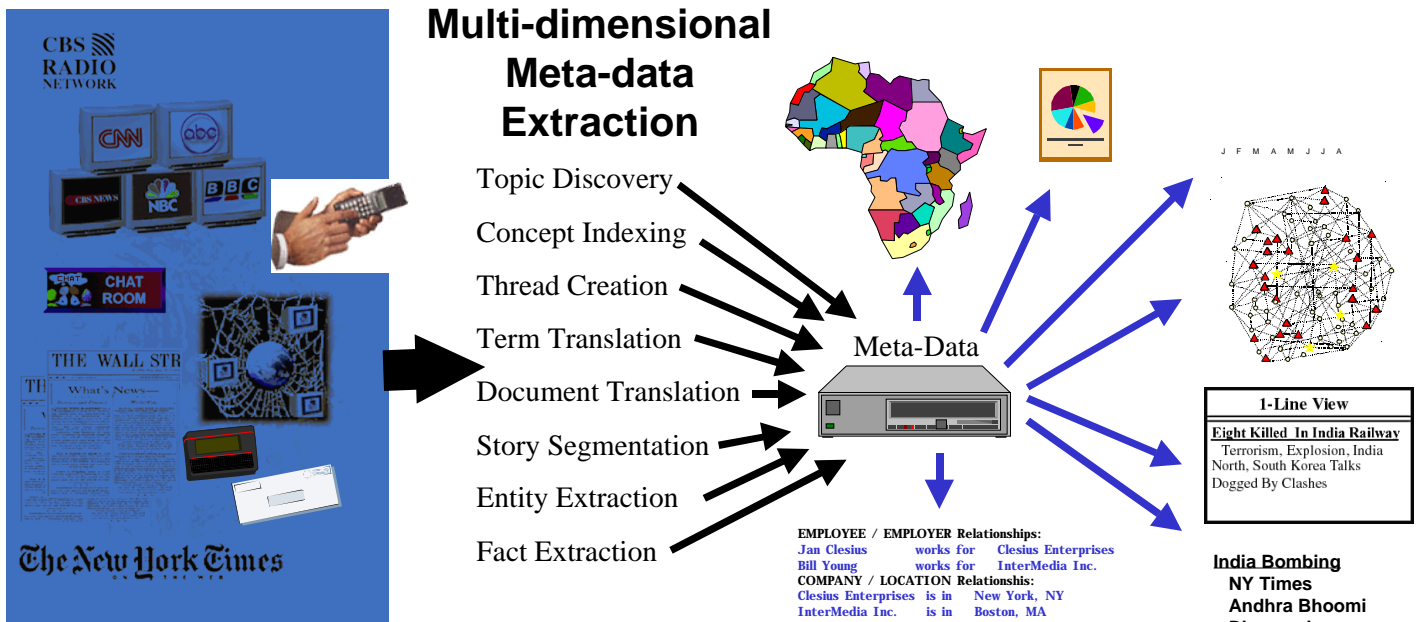

Statistical Models of Text: From Bags of Words to Structure

Ralph Weischedel

17 April 2000



Extraction Vision

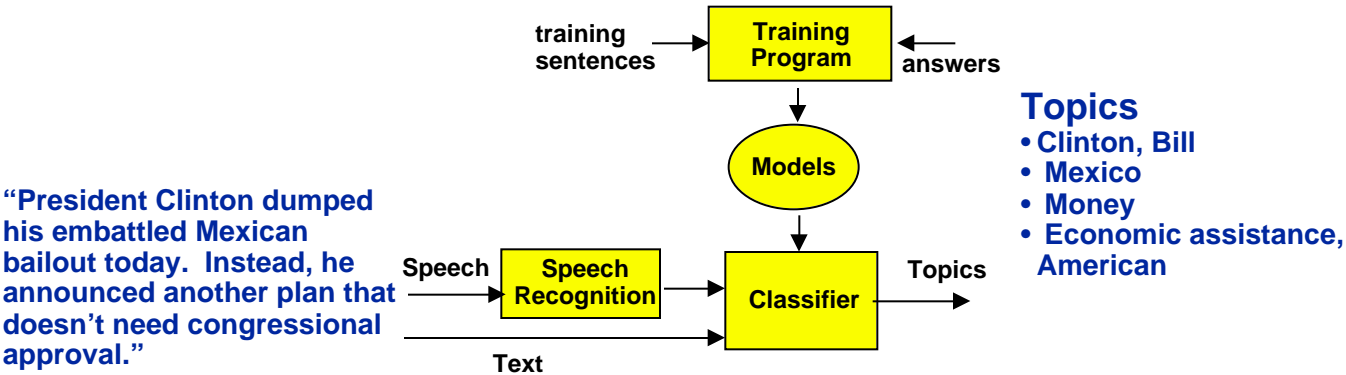


Outline

Statistical models that support feature extraction

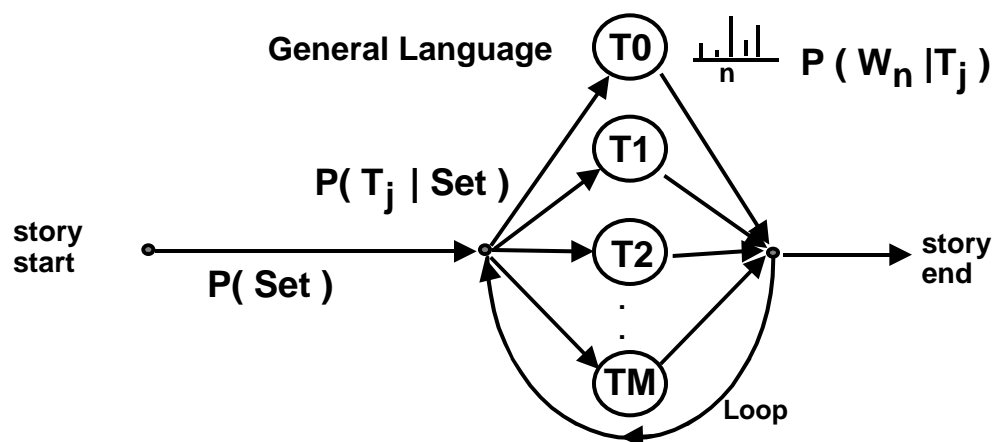
- **Bags of words**
 - Topic extraction
- **Sequences (HMMs)**
 - Name extraction and classification
- **Lexicalized probabilistic context-free grammars**
 - Parses
 - Facts/relationships
- **TBD**
 - Propositions

Topic Extraction via Bag of Words



Generative Model of Story and Topics

- First, choose a Set of topics, $T_0 \dots T_M$
- For each word in story:
 - Choose a topic according to $P(T_j | \text{Set})$
 - Choose a word according to output distribution $P(W_n | T_j)$
 - Loop



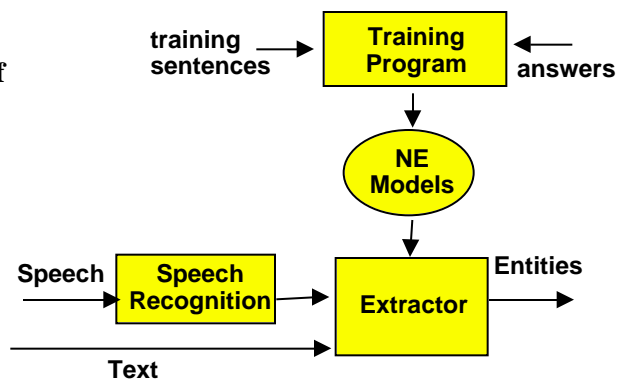
Topic Classification on Broadcast News

- Trained on 1 year of stories from July '95 to Jun '96 (42,502 stories)
- Tested on 989 stories from July '96
- Allowed 4,627 topics that occur at least twice
- OOT (out-of-topic) rate was 2.45%
- **Results:**
 - 75.8% of the first choice topics are among the annotated labels
 - 63.6% for a simple likelihood-based method
 - 45% for the traditional tfidf measure used in IR
- **On cursory examination of errors, often the recognized topic was correct and the annotator failed to include it.**



Name Extraction via HMMs

The delegation, which included the commander of the U.N. troops in Bosnia, Lt. Gen. Sir Michael Rose, went to the Serb stronghold of Pale, near Sarajevo, for talks with Bosnian Serb leader Radovan Karadzic.



The delegation, which included the commander of the U.N. troops in Bosnia, Lt. Gen. Sir Michael Rose, went to the Serb stronghold of Pale, near Sarajevo, for talks with Bosnian Serb leader Radovan Karadzic.

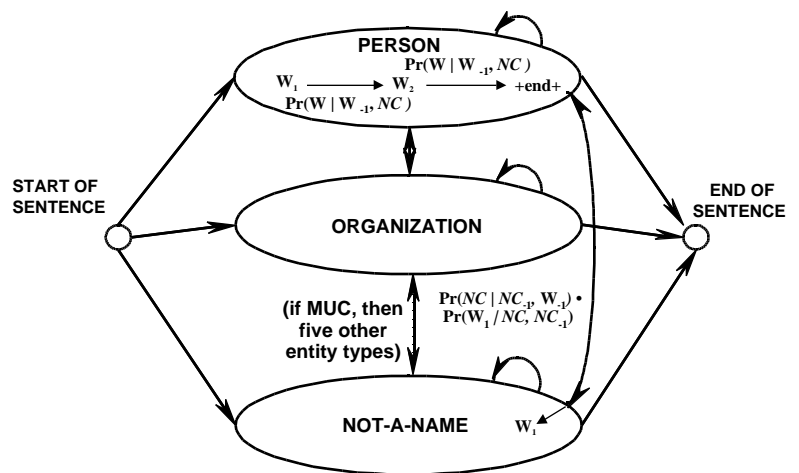
- Locations
- Persons
- Organizations

- Prior to 1997 - no learning approach competitive with hand-built rule systems
- Since 1997 - Statistical approaches (BBN, NYU, MITRE) achieve state-of-the-art performance

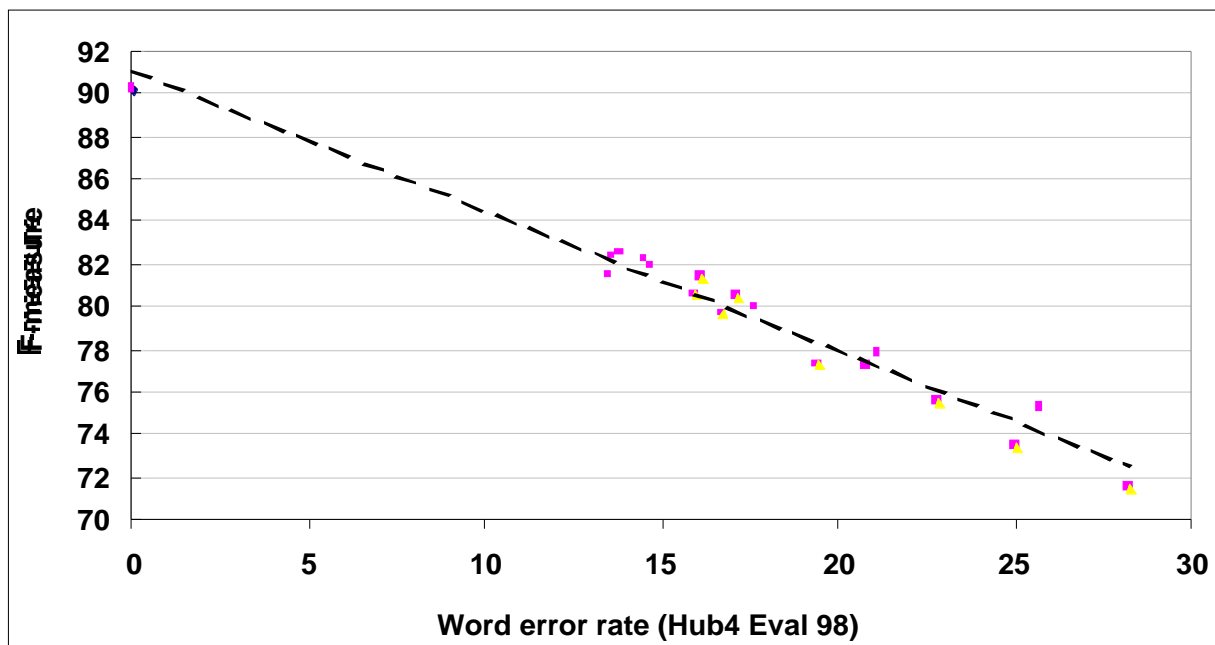
A Hidden Markov Model

Structure of Model

- One language model for each category plus one for other (not-a-name)
- The number of categories is learned from training
- Bi-gram transition probabilities



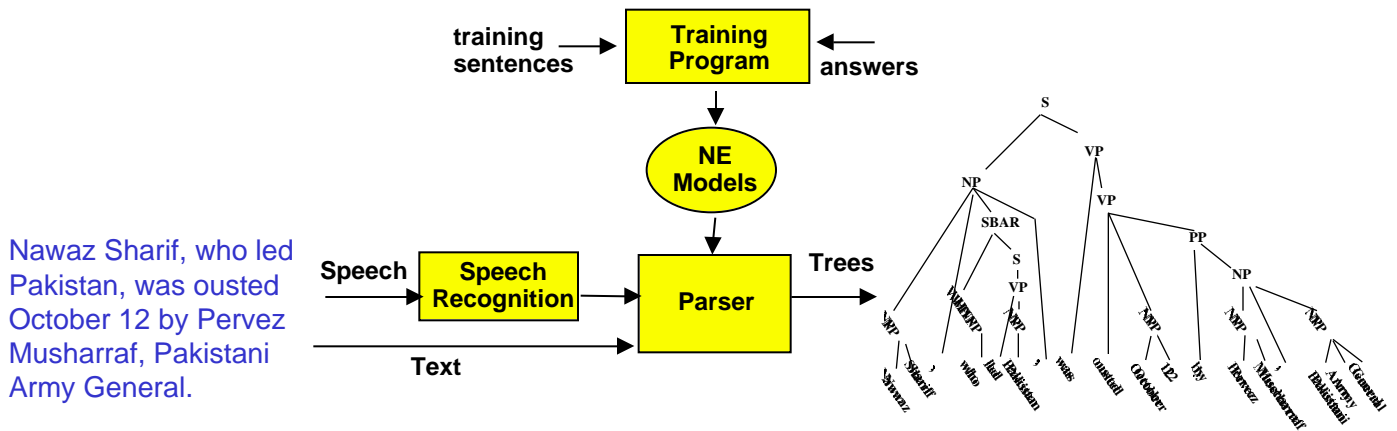
Effect of Speech Recognition Error



BBN and NIST found Identifinder performance degrades 0.7 points of F per 1% WER



Parsing via Lexicalized Probabilistic CFGs

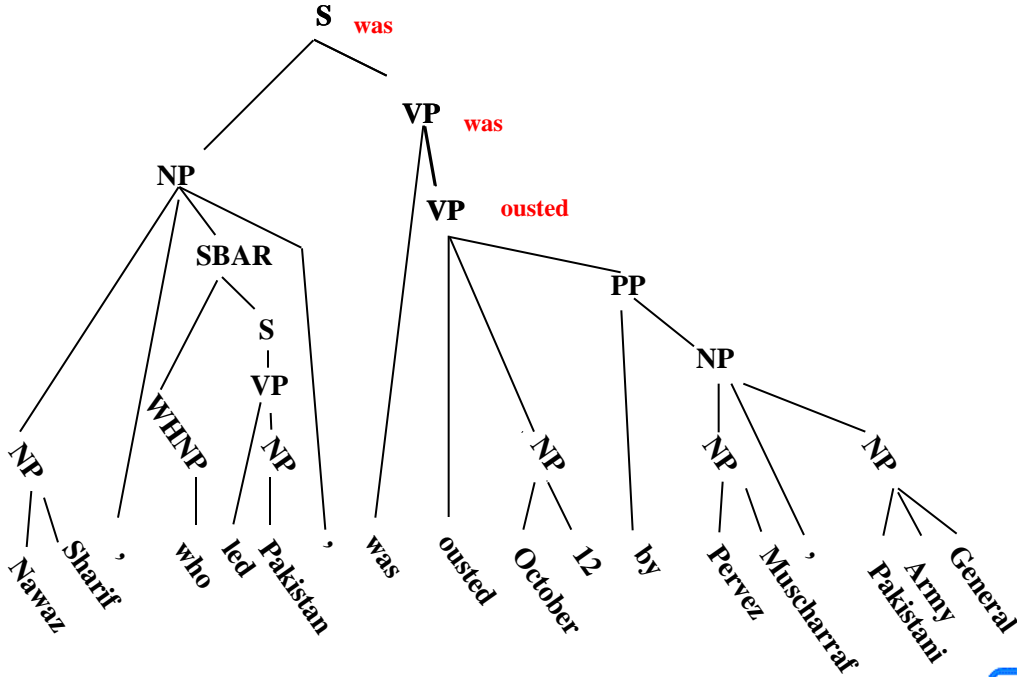


Nawaz Sharif, who led Pakistan, was ousted October 12 by Pervez Musharraf, Pakistani Army General.

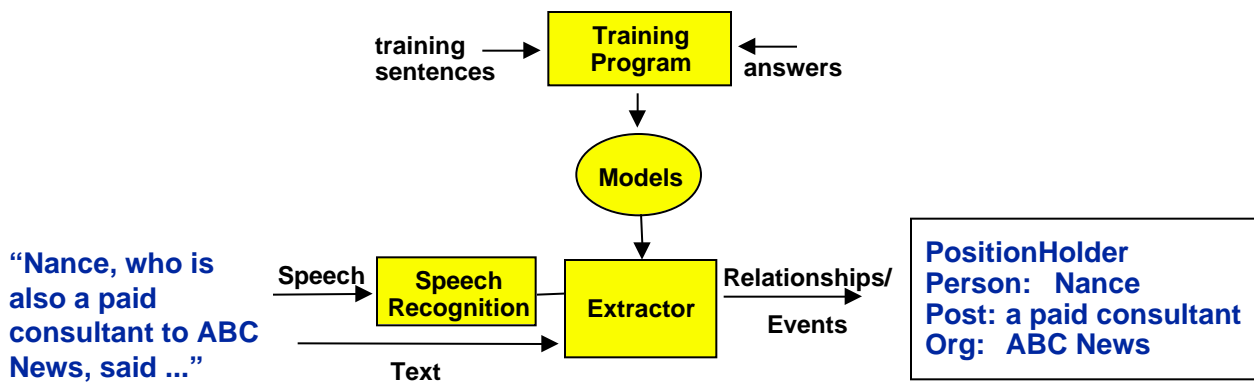
- **Prior to 1990 - accuracy for non-statistical parsers around 65%**
- **Since 1995 - Statistical parsers (IBM, UPenn, Brown and BBN) achieve 85-90% accuracy**



Example of Generating a Parse Tree

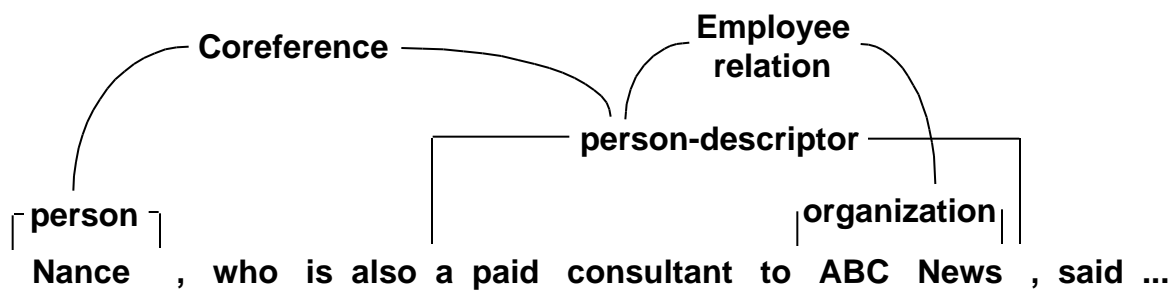


Extracting Facts via LPCFG



- 1998 - First state-of-the-art trainable system (70% accuracy)

Type of Annotation Required



- **Training data consists ONLY of**
 - Named entities (as in NE)
 - Descriptor phrases (for TE)
 - Descriptor references (for TE)
 - Relation/events to be extracted (for TR)

The Sentential Model

$$p(M|W) = \frac{p(M,W)}{p(W)} = \max_T \frac{p(M,T,W)}{p(W)}$$

- **Search Criterion: find M such that p(M | W) is maximized**
- **Since p(W) is constant, search for:**

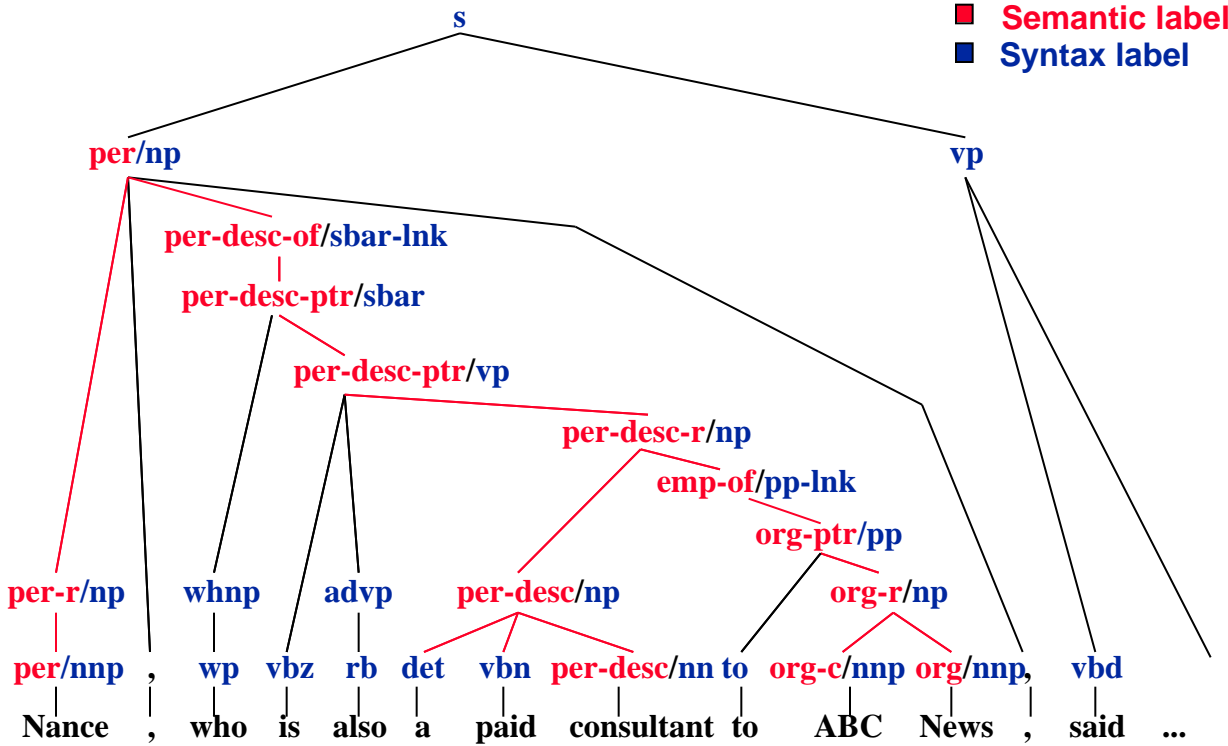
$$\max_M p(M, T, W)$$

- **Model the probability as the product of the probabilities of generating each element in the tree**

$$p(M, T, W) = \prod_{e \text{ tree}} p(e | h)$$

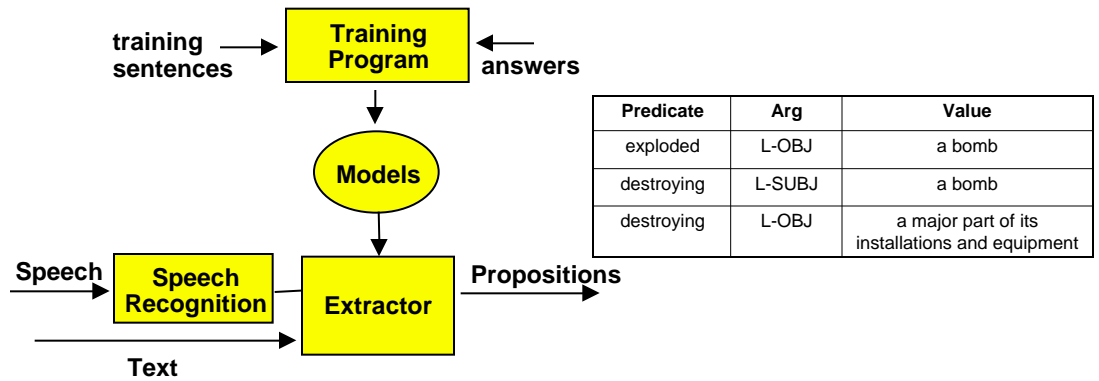


Augmented Semantic Tree



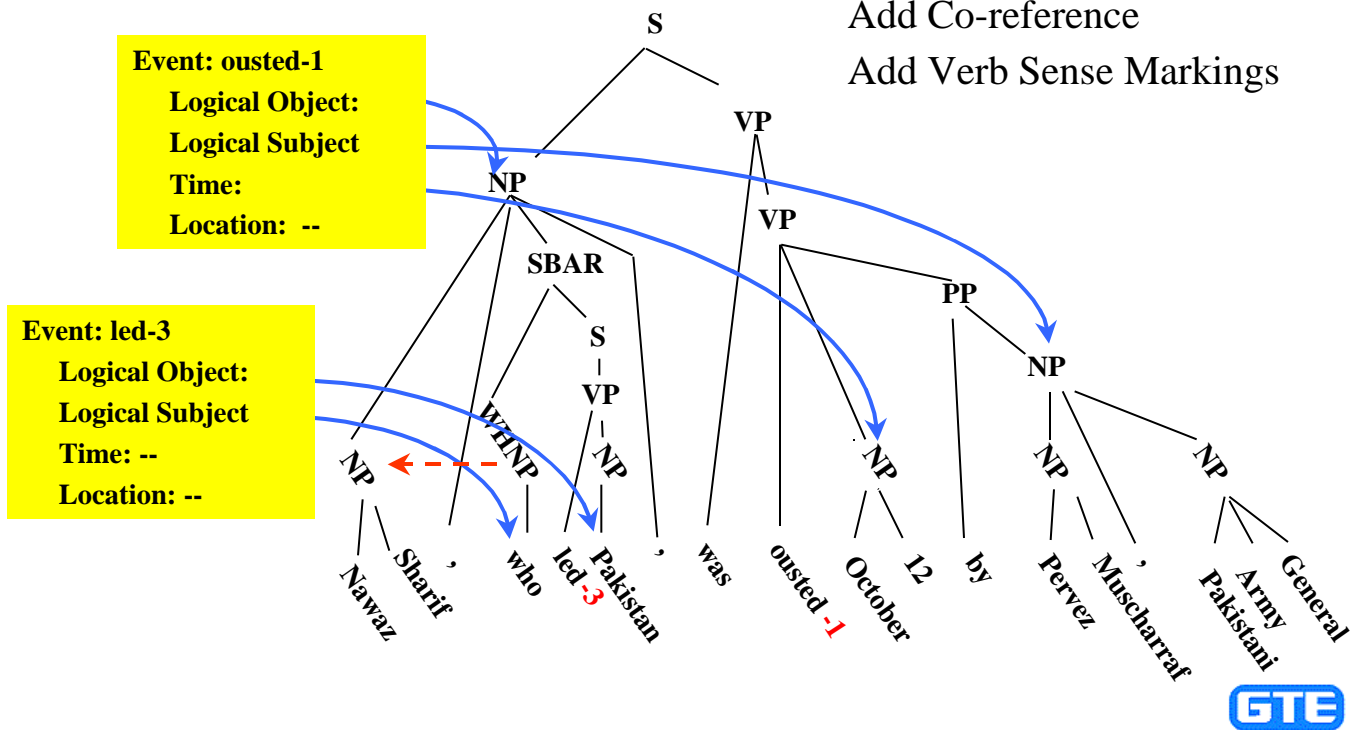
Propositions via TBD

Within the past two months, a bomb exploded in the offices of the El Espectador in Bogata, destroying a major part of its installations and equipment.

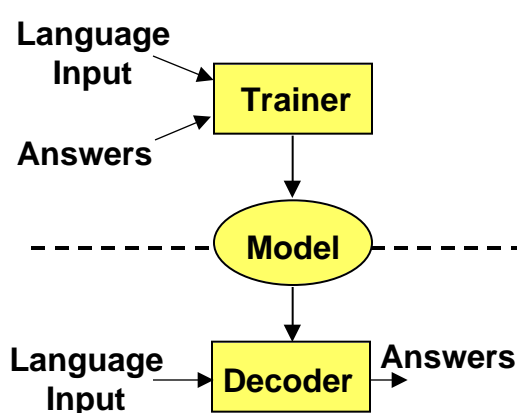


Towards a Proposition Bank

- Add Predicate/Argument Markings
- Add Co-reference
- Add Verb Sense Markings



Statistical Speech/Language Modeling



Technology	Input	Answers
• Speech recognition	audio	transcription
• OCR	image	characters
• Speech understanding	audio	response
• Topic classification	document	topics
• Topic detection	text/speech	clusters
• Topic tracking	text/speech	relevant stories
• Story segmentation	speech	stories
• Information retrieval	query	text/speech
• Named entity extraction	text/speech	names & types

Advantages

- **Mathematically rigorous approach**
- **State-of-the-art performance**
- **Highly robust in the face of degraded input**
- **Language independent, requiring only annotated training data**
- **Affordable annotation**
 - Only domain knowledge is needed
 - Can be performed by students/interns

