

(Statistical) Natural Language Processing

Murat Saraçlar

(Statistical) Speech and Language Processing

Murat Saraçlar

Speech and Language Processing

- Ultimate Goal: Interacting with computers using human languages (as opposed to computer languages)
 - Talk to the computer (HAL, Star Trek)
 - Chat with it ?
- Hard and interesting problem, also a very real problem
- Not Linguistics

What do we need?

- Apart from being Artificially Intelligent,
- Speech Recognition (and lip reading)
 - Natural Language Understanding
 - Natural Language Generation
 - Speech Synthesis
 - Information Retrieval
 - Information Extraction

Speech and Language Processing

- Natural Language Processing
- Computational Linguistics
- Speech Recognition
- Speech Synthesis

Properties of Natural Language

- What do we know about language?
- Ambiguity
 - Source of humor
 - Difficult to be precise
 - Can easily be misunderstood, misinterpreted
- Hierarchical Structure
 - Different levels of representation

Issues in NLP

- Complexity/Efficiency
- Robustness: Dealing with humans
- Specificity/Generality
- Knowledge/Data
- Inference: very difficult
- Evaluation on real data (corpora=derlem)

Some NLP Applications

- Dialogue, Question Answering
- Machine Translation
- Information Extraction / Summarization
- Information Retrieval / Search Engines
- Word processing: spelling correction, grammar checking
- Intelligent (and more natural) User Interfaces

Levels of Language

- **Phonetics and phonology:** The study of linguistic sounds that make up the words.
- **Morphology:** Meaningful components of words and how they come together.
- **Syntax:** How words come together.
- **Semantics:** Meaning
 - lexical and compositional
- **Pragmatics:** Accomplishing goals.
- **Discourse:** How sentences come together.

Resolving Ambiguity

I made her duck

- I cooked duck for her
- I cooked the duck belonging to her
- I created the (plastic) duck she owns
- I caused her to duck
- I turned her into a duck

Resolving Ambiguity

I made her duck

- **Phonetics and phonology:** Eye/I, made/maid
- **Morphology:** duck (Verb/Noun), her (dat/poss)
- **Syntax:** make (one or two objects)
 - (I (made (her) (duck))
 - (I (made (her duck)))
- **Semantics:** make (create/cook)
- **Pragmatics:**
- **Discourse:**

Models and Algorithms

- (Weighted) (Finite) State Machines
 - Automata, transducers
 - (hidden) Markov models
- Formal Languages and rules systems
 - Regular, context-free
- Graph Search: dynamic programming, A*
- Logic
- Probability Theory and Machine Learning

Demo

- Zemberek
 - Spelling
 - Analysis
 - ASCII -> TR
 - Syllabification
 - Language ID

Chomsky Hierarchy

Language	Mechanisms	Examples
Regular	Regular Expressions (Weighted) Finite-State Automata and Transducers	xa^ny Phonology Morphology Tagging
Context-Free	Context-Free Grammars Push-down Automata	a^nb^n Syntax
Context-Sensitive	Unification Grammars Tree-Adjoining Grammar	$a^nb^mc^nd^m$

Chomsky Hierarchy: Rule Types

Type	Name	Rule Type
0	Recursively enumerable Turing Equivalent	$\alpha \Rightarrow \beta$
1	Context Sensitive	$\alpha A \beta \Rightarrow \alpha \gamma \beta$
2	Context Free	$A \Rightarrow \gamma$
3	Regular	$A \Rightarrow xB \quad A \Rightarrow x$

NLP Perspective

- Whether language is really context-free or not is irrelevant
 - Want to get 90-95 percent correct. Forget the rare ones.
 - Go from linear finite-state to polynomial (n^3) CF
 - to n^6 for many context-sensitive formalisms
- Use of statistical techniques enables aggressive and effective pruning

EE586

- Coordinator: Murat Saraçlar (<http://busim.ee.boun.edu.tr/~murat/>)
- Guest Lectures by:
 - Kemal Oflazer (Sabanci U)
 - Tunga Gungör (CmpE)
 - Cem Say (CmpE)
- Hands on and interactive

More Info

- **Textbook:** Daniel Jurafsky and James H. Martin, [Speech and Language Processing](#), Prentice Hall, 2000.
- **Reference Text:** Chris Manning and Hinrich Schütze, [Foundations of Statistical Natural Language Processing](#), MIT Press, 1999.
- **Prerequisites by topic:** Probability, Programming

Course Coverage

- Subproblems:
 - Disambiguation and Annotation
 - Tagging, Parsing, Language Modeling
- Tools:
 - Formal: (Weighted) Finite-State Automata and Transducers, Context-Free Grammars
 - Statistical Models: N-gram LMs, HMMs, PCFGs
 - Statistical Methods: Smoothing, EM, MaxEnt...

Topics

- | | |
|--|---------------------------------|
| 1. Regular Expressions and Automata | 9. Semantics |
| 2. Morphology and Finite-State Transducers | 10. Machine Translation |
| 3. Computational Phonology | 11. Dialog |
| 4. N-gram Language Models | 12. Natural Language Generation |
| 5. Hidden Markov Models | |
| 6. Part-of-Speech Tagging | |
| 7. Context Free Grammars | |
| 8. Probabilistic Parsing | |

(Statistical) Speech and Language Processing

Murat Saraçlar

Boğaziçi University
EE Department

Demo: Broadcast News Transcription

- Towards Automatic Closed Captioning
- Real-Time Transcription from LIVE TV
- Low Latency (typically <1sec)
- Very Large Vocabulary (>200K words)
- High Word Accuracy (>80%)
 - Wolf Blitzer (CNN): 96%
 - Henry Kissinger: 64%

Projects

- Broadcast News Transcription and Retrieval
- Speech to Speech Machine Translation
- Computer Aided Transcription
- Computer Aided Pronunciation Training
- Reading Tutor
- Sign Language Processing

Statistical Formulation of Automatic Speech Recognition

$$\hat{W} = \arg \max_w P(W | A)$$

- A: acoustic signal (sequence of vectors)
- W: sentences (sequence of words)

Statistical Formulation of Automatic Speech Recognition

$$\hat{W} = \arg \max_W P(A|W)P(W)$$

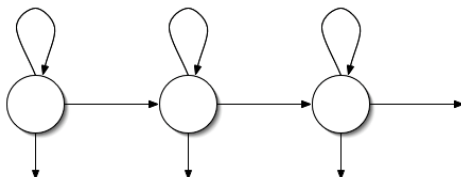
- $P(A|W)$: the acoustic model (channel)
- $P(W)$: the language model (source)
- $\arg \max$: hypothesis search (decoding)

The Front-End

- Sound waves converted to a digital signal (16KHz, 16bits linear or 8KHz, 8bits μ -law)
- Short-time analysis of the frequency content.
- Mel-Frequency Cepstral Coefficients (MFCC)
 - Mel-Frequency: log-spaced filter banks.
 - Cepstrum: computed as DCT of log Energy
- Dynamic info added by time-derivatives
- A 39-dimensional vector every 10ms

The Acoustic Model

- Words represented as concatenation of phones (or phones in context: triphones)
- Each (tri)phone is modeled by a Hidden Markov Model
 - A finite state Markov chain with outputs



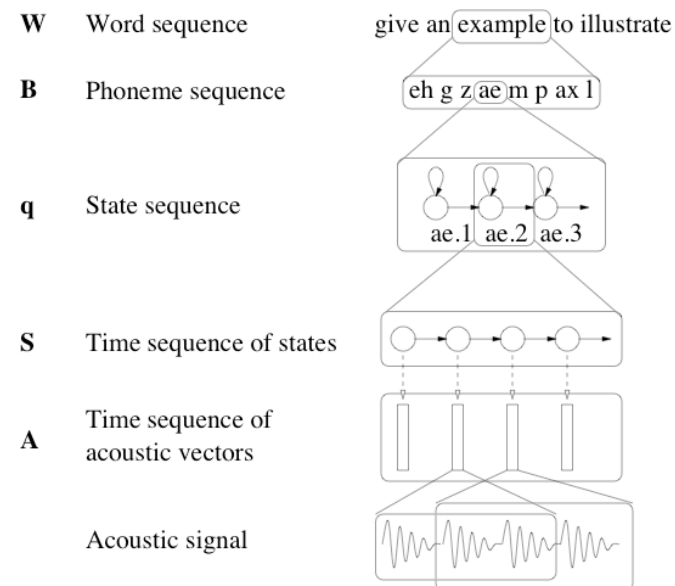
The Language Model

- Words modeled as a sequence
- N-gram models based on a finite history (Markov assumption)
- Typically trigram (surprisingly good):

$$P(W) = \prod_{i=1}^N P(w_i | w_{i-2}, w_{i-1})$$

Hypothesis Search

- The search space can be organized into a finite state network.
- The network can be optimized (determinized, minimized) using weighted finite state automata (transducers) theory.
- Viterbi decoding (time-synchronous beam search).



State-of-the-art ASR Performance

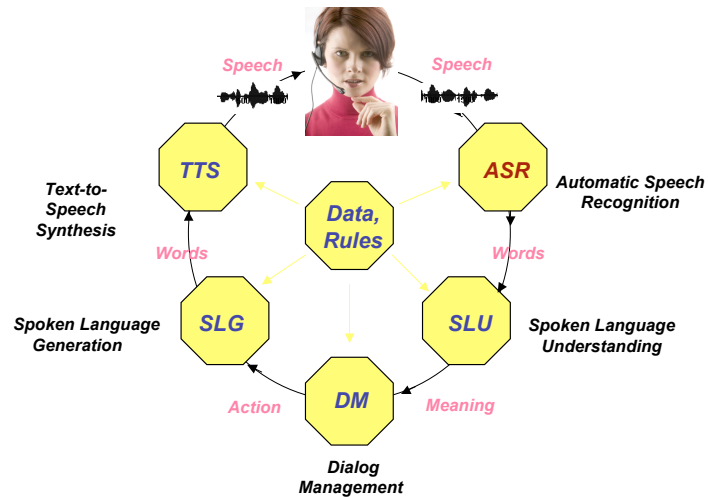
Task	TYPE OF SPEECH	VOCABULARY SIZE	WORD ERROR RATE
Connected Digit Strings	Read Text	10	0.2% RT
Connected Digit Strings	H-M Real-time Conversational Telephone	10	2% RT
Wall Street Journal	Read Text	500,000	5 %
Broadcast News	Mixed	210,000	10% (18%) RT
Switchboard	H-H Conversational Telephone	28,000	16% (24%) RT
Natural Language Dialog	H-M Conversational Telephone	8000	20% RT

10-15% relative reduction per year.

Applications of ASR

- Dictation
 - Medical transcription
- Customer Service Applications
 - Directed dialog / Form filling
 - Natural language dialog
- Retrieval for Spoken Communications
 - News archives, Lectures, Meetings
 - Teleconferences, phone conversations
 - Voice mail
- Speech-to-Speech Translation

The Speech Technology Chain



Spoken Language Understanding

“I would like to make a collect call”

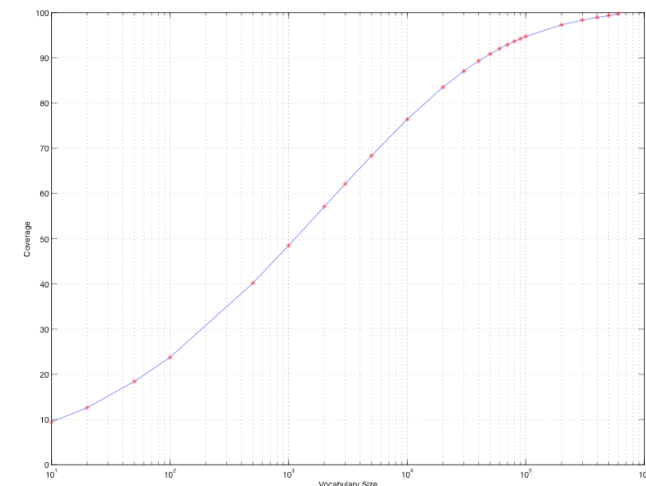
“Show me the flights from Istanbul to Ankara on Monday”

- Call Classification
 - Utterance Classification
 - Dialog Management
- Information Extraction
- Natural Language Processing
 - Part-of-Speech Tagging
 - Parsing

Large Vocabulary ASR for Turkish

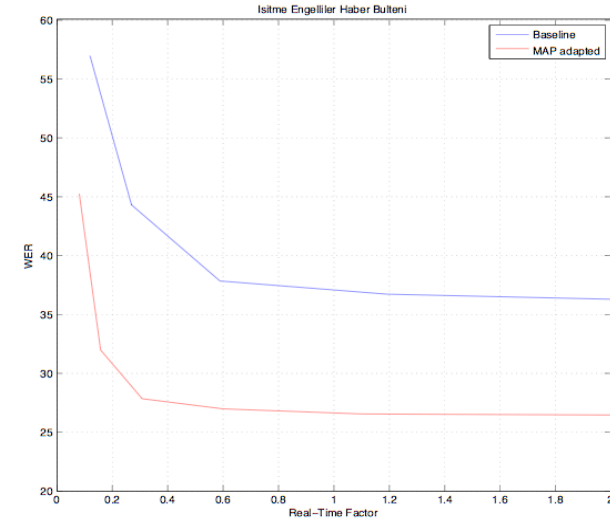
- Agglutinative Language
Uygar+laş+tır+a+ma+ya+cak+lar+ımız+dan+mış+sınız+casına
- Free Word Order
 - Adam kitabı okudu.
 - Adam okudu kitabı.
 - Kitabı adam okudu.
 - Kitabı okudu adam.
 - Okudu kitabı adam.
 - Okudu adam kitabı.

Vocabulary Coverage for Turkish

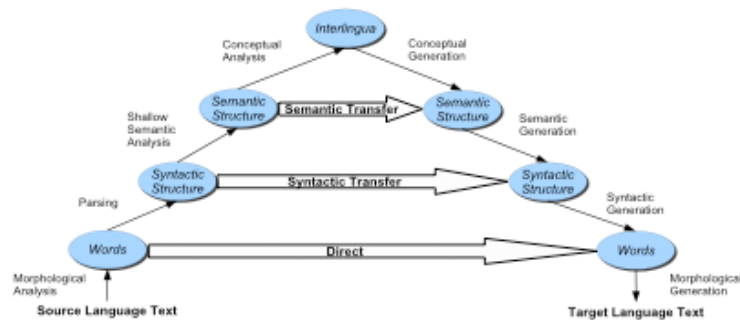


Turkish Broadcast News Transcription Demo

- 50K word vocabulary (OOV=8.9%)
- Acoustic models trained on read speech and adapted
- Language model trained on books, news, text collected from the web
- Using the AT&T Decoder
- Tests on TRT2 İşitme Engelliler Haber Bülteni



Machine Translation



Statistical Formulation of Machine Translation

$$\hat{T} = \arg \max_T \text{faithfulness}(T, S) \text{fluency}(T)$$

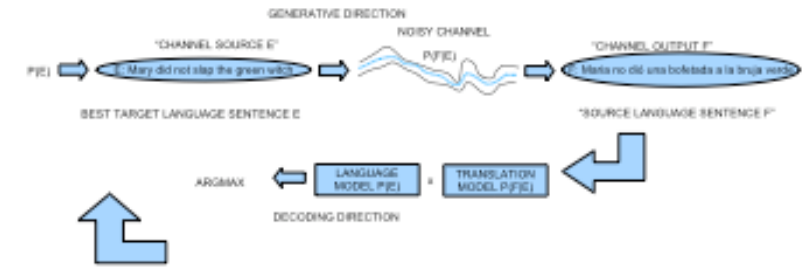
- S: source language
- T: target language

Statistical Formulation of Machine Translation

$$\hat{T} = \arg \max_T P(S|T)P(T)$$

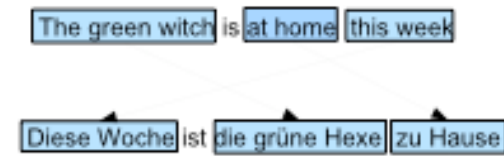
- $P(S|T)$: the translation model (channel)
- $P(T)$: the language model (source)
- $\arg \max$: hypothesis search (decoding)

Noisy Channel Model for Machine Translation



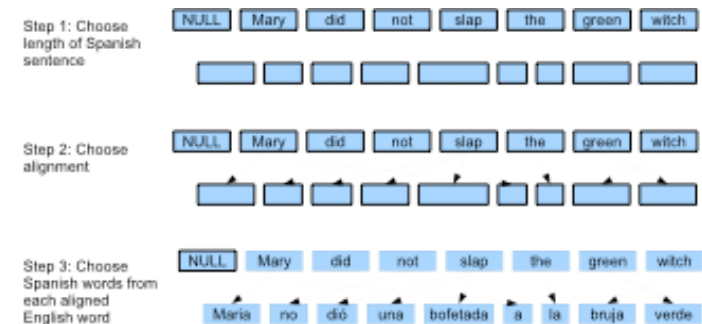
The Phrase Based Translation Model

- $S=s_1, \dots, s_n$
- $T=t_1, \dots, t_n$
- $d()$: distortion



$$P(S|T) = \prod_{i=1}^n \phi(\bar{s}_i, \bar{t}_i) d(start(\bar{s}_i | \bar{t}_i) - end(\bar{s}_{i-1} | \bar{t}_{i-1}))$$

Alignment in Machine Translation IBM Model I



EE586: Natural Language Processing

<http://busim.ee.boun.edu.tr/~murat/teaching/EE586>

- | | |
|--|---------------------------------|
| 1. Regular Expressions and Automata | 9. Semantics |
| 2. Morphology and Finite-State Transducers | 10. Machine Translation |
| 3. Computational Phonology | 11. Dialog |
| 4. N-gram Language Models | 12. Natural Language Generation |
| 5. Hidden Markov Models | |
| 6. Part-of-Speech Tagging | |
| 7. Context Free Grammars | |
| 8. Probabilistic Parsing | |