Natural Language Processing: A Brief Review

Eduard Hovy

Information Sciences Institute
University of Southern California
www.isi.edu/~hovy

USC INFORMATION SCIENCES INSTITUTE

ISI Information Sciences Institute

What is NLP?

- Machine Translation (MT)
- Speech Recognition (ASR)
- Information Retrieval (IR)
- Information Extraction (IE)
- Text Summarization

Phase 1: Getting started

1950–65

The Grand Challenge: **MT**

- Warren Weaver: memorandum, 1946
- MT demo: IBM/Georgetown U, 1954 (USAF)
- Journal *Machine Translation*, 1954 … later *Computational Linguistics*
- International MT conferences: 1954, 1956, 1958
  - at 1958 conference: MT/NLP ↔ IR
  - Luhn: auto-summaries of papers in one session
- Very limited computer space/power: 7 minutes to parse long sentence
- Tried both statistical and symbolic methods
- ALPAC report, 1964

IR: for the librarians

- Intense manual effort to build index structures
- Cleverdon: the Cranfield aeronautics text evaluation experiments

Phase 2: Trying theory

1965–75

- **NLP**
  - syntax: Transformational Grammar, then other approaches
  - lexicon efforts: polysemy, etc.
  - processing: rather ad hoc, then finite state automata (Woods et al.)
- **IR**
  - lots of work on indexing books and articles
  - start of vector spaces: Salton at Cornell
  - system construction: intense manual effort
- **Speech**
  - units: single words
  - system construction: intense manual effort to model articulatory channel
- Pre-computational **semantics**: Masterman, Ceccato

Phase 3: Higher ambitions!

1975–85

- **NLP**
  - formal and informal semantics: Situation Semantics (Barwise and Perry ~77), DRT (Kamp 80); Frames (Minsky 75), Semantic Nets (Bobrow and Collins 75), Conceptual Dependency etc. (Schank 77–85; Jackendoff 80; Sowa 80s)…
  - processing: ATNs (e.g., LUNAR, Woods 78)
- **AI**
  - SHRDLU (Winograd 73) and TALE-SPIN (Meehan 75)
- **IR**
  - vector spaces firmly established
  - system construction: automatic, with some tuning
- **Speech**
  - triumphant introduction of learning methods: HMMs at CMU (Baker)
  - system construction: some learning, and tuning
  - units: phrases

Phase 4: Two methodologies

1985–95

- **NLP**: **theoretical** side
  - logical form and well-formed formulas
  - formal grammars: HPSG, GPSG and all the other PSGs
  - processing: unification as the Great Answer (Shieber 86)
- **MT**
  - statistical MT (Brown et al. 90s); the Statistics Wars

theory-driven

- **NLP**: **practical** side
  - IE (MUC competitions)
  - preprocessing, alignment, etc. tools (Church, Brill, etc.)
  - Penn Treebank and WordNet
- **IR**
  - TREC competitions (1990–); various tracks
  - moving to the web
- **Speech**
  - system construction: learning HMMs (bi-, trigrams)
  - simple dialogues (ATIS)
  - DARPA evaluations and systems

experiment-driven

Phase 5: Statistics 'wins'

1995–05

- **NLP**
  - machine learning of (almost) everything; statistics-based parsing (Collins, Charniak, Hermjakob)
  - large networks, centers, and corpora of all kinds (ELSNET, Penn Framebank, etc.); LREC, EMNLP, and Very Large Corpora conferences
  - shallow semantics: WordNet 1.6 (Miller, Fellbaum) and the other Nets
  - practical applications: summarization
- **IR**
  - mathematical formulation of theories in vector spaces and language models
  - ever larger scope: web, cross-language IR, rapid classification…
  - QA
- **MT**
  - statistical MT tools (Knight et al.) and automated MT evaluation (Papineni et al.)
- **Speech**
  - mathematical formulation of theories and machine learning of more than just HMMs
  - dialogue: adding some context; linking with NLG and synthesis (Verbmobil, DARPA Communicator projects)
  - toward unlimited vocabulary and noisy backgrounds

Phase 6: Today

2005–

Where are we today?

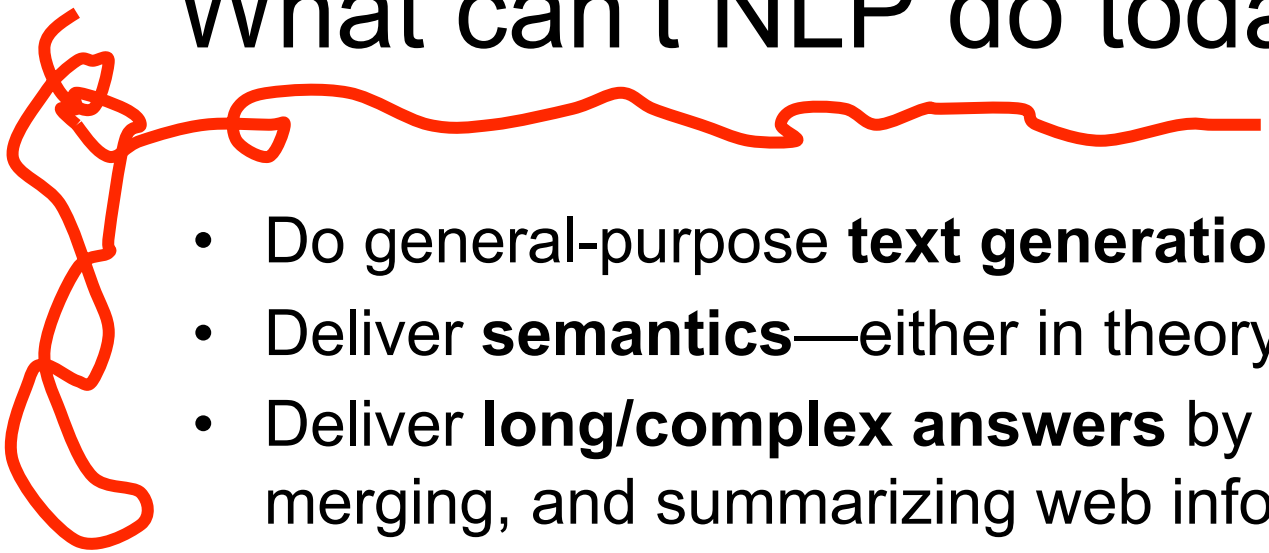
Technology competitions

- TREC
  - Started at NIST in 1991
  - IR, Web retrieval, Interactive IR, Filtering, Video retrieval, CLIR, QA…
  - Highly successful formula—push research and share results
- CLEF
  - Copy of TREC in Europe
  - Latest CLEF CLQA, others
- NTCIR
  - Like TREC, held in Tokyo last week
  - IR, CLIR, QA, Summarization…
  - Organizers:
    - Organizing Chair: Jun Adachi, NII
    - Program Chair: Noriko Kando, NII
- Others:
  - ACE, DUC…

NLP in the world

- There are between 10,000 and 15,000 NLP practitioners in the world:
  - ISCA—3000 members?
  - ACL—2000 members
  - SIGIR—1000 members
  - IAMT—400 members
- There are over 20 conference series: ICSLP, ACL (+ NAACL-HLT, EACL), COLING, LREC, SIGIR, EMNLP, MT Summit (+ AMTA, EAMT, AAMT), RANLP, PACLING, INLG, ROCLING, TMI, CICLing… plus numerous workshop series

What can't NLP do today?

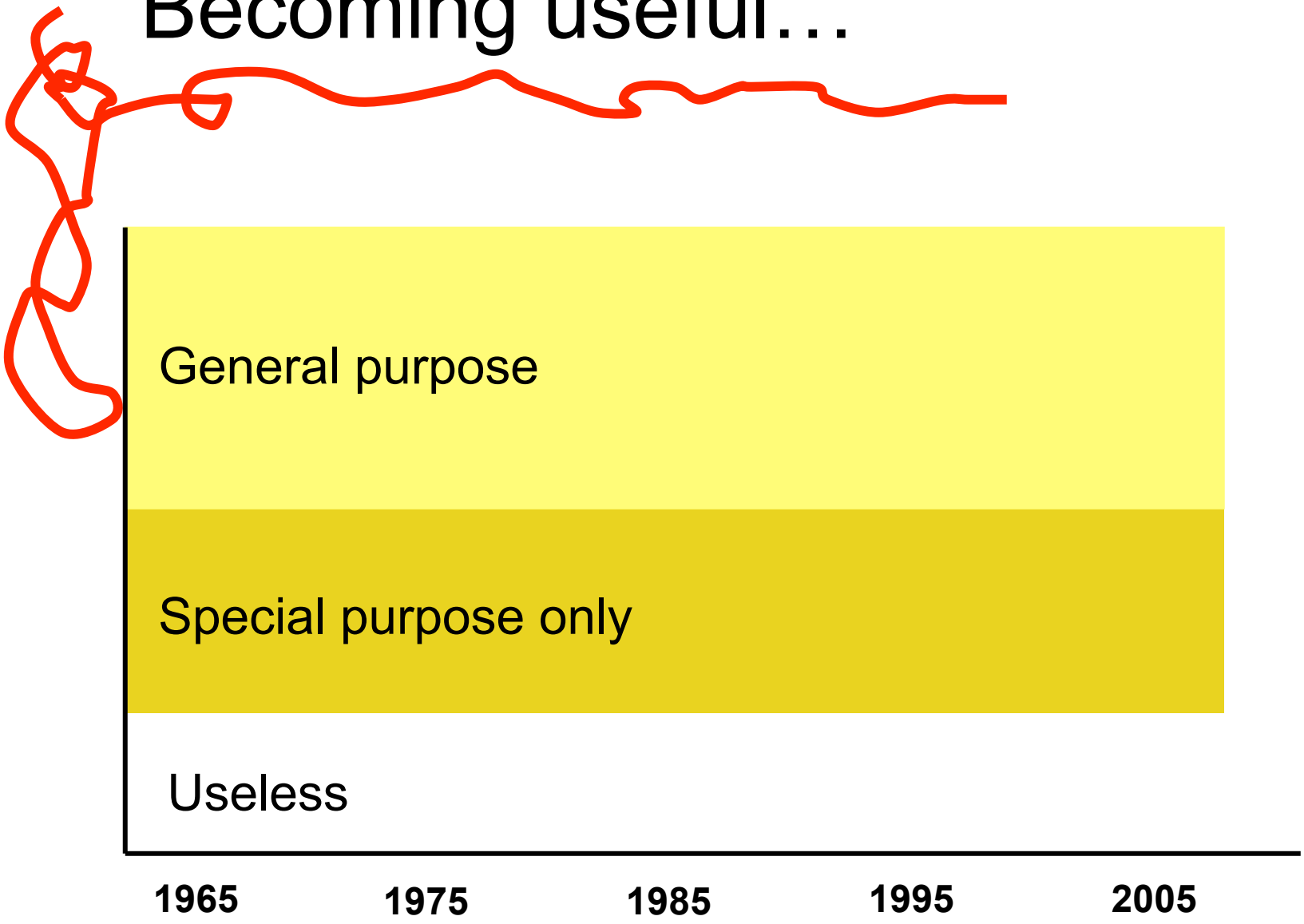


- Do general-purpose **text generation**
- Deliver **semantics**—either in theory or in practice
- Deliver **long/complex answers** by extracting, merging, and summarizing web info
- Handle extended **dialogues**
- **Read and learn** (extend own knowledge)
- Use **pragmatics** (style, emotion, user profile…)
- Provide significant contributions to a **theory of Language** (in Linguistics or Neurolinguistics) or of **Information** (in Signal Processing)
- etc.…

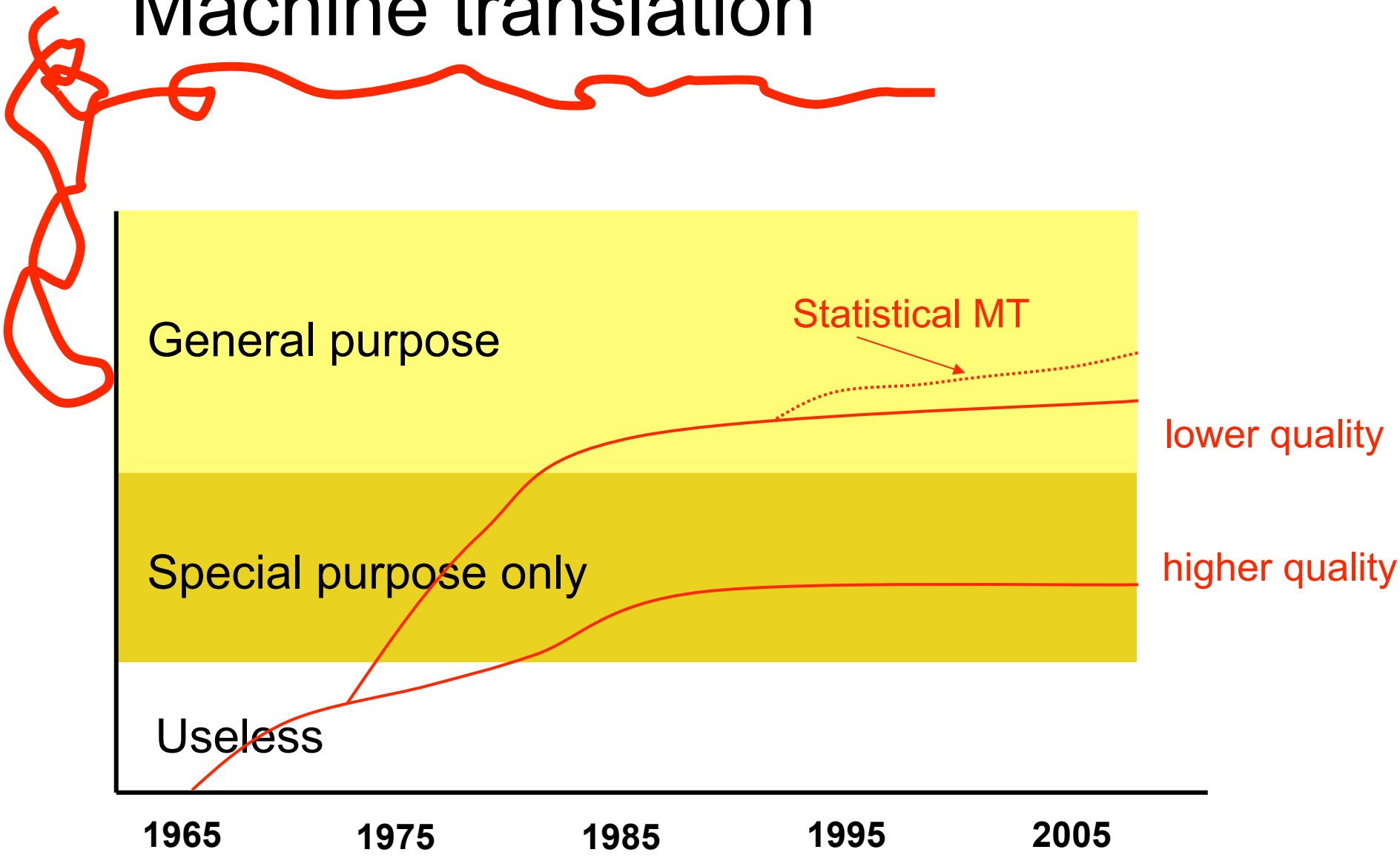
What can NLP do (robustly) today?

- Reliable surface-level **preprocessing** (POS tagging, word segmentation, NE extraction, etc.): 94%+ — 90s–
- Shallow syntactic **parsing**: 92%+ for English (Charniak, Collins, Lin) and deeper **analysis** (Hermjakob) — 00s–
- **IE**: ~40% for well-behaved topics (MUC, ACE) — 80s–
- **Speech**: ~80% large vocab; 20%+ open vocab, noisy input — 80–90s
- **IR**: 40% (TREC) — 80–90s
- **MT**: ~70% depending on what you measure — 80s–
- **Summarization**: ? (~60% for extracts; DUC) — 90–00s
- **QA**: ? (~60% for factoids; TREC) — 00s–

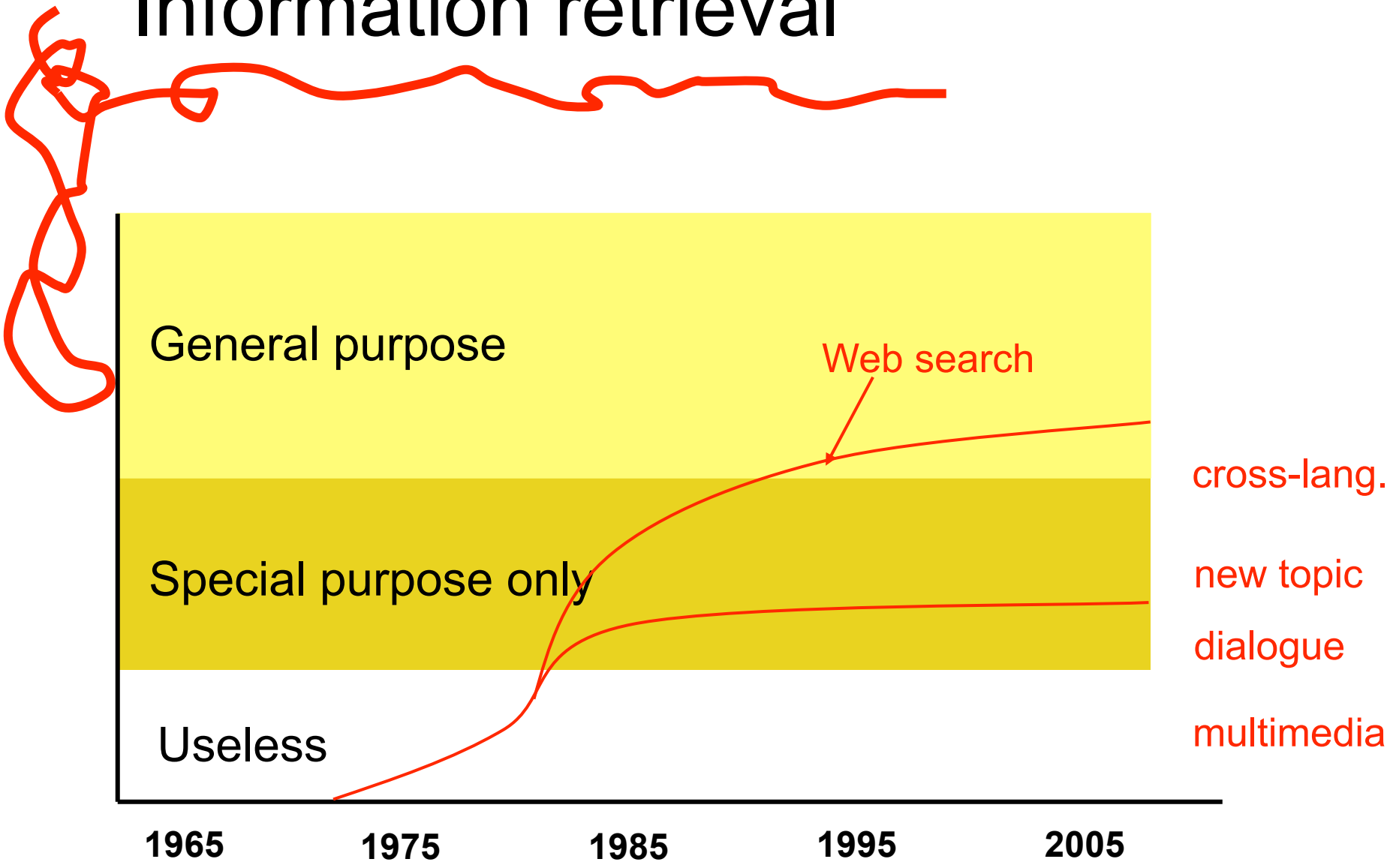
Becoming useful…



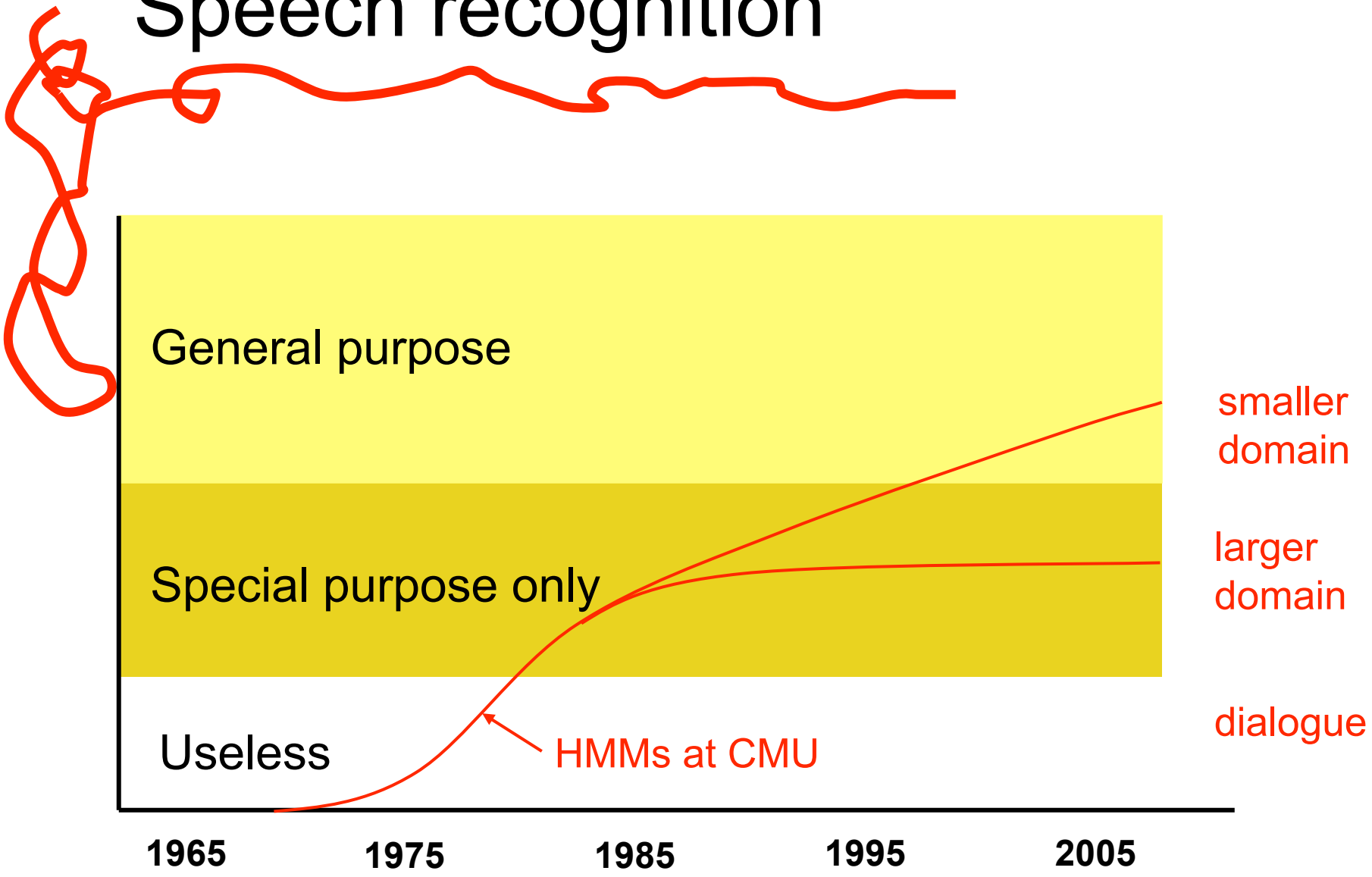
Machine translation



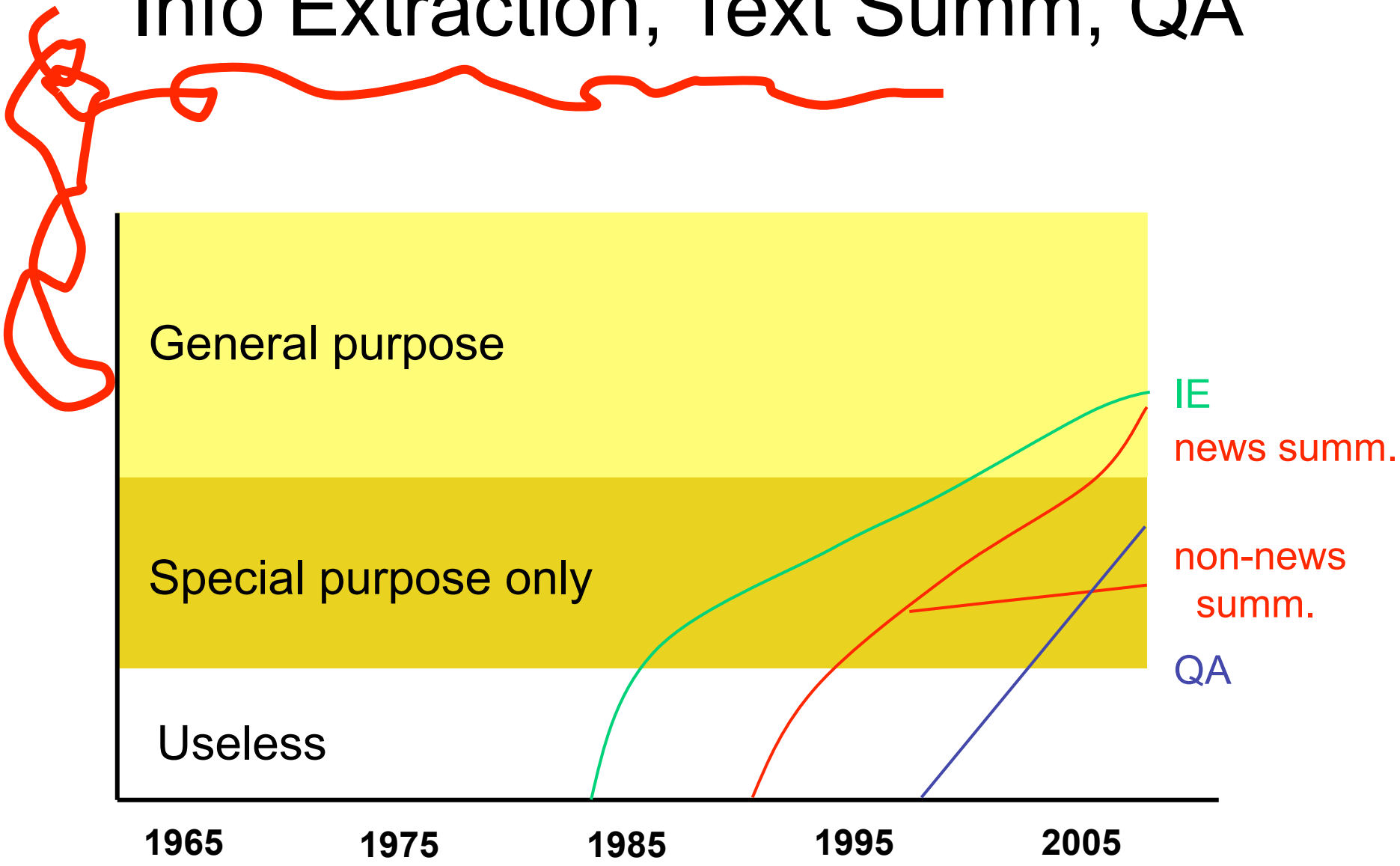
Information retrieval

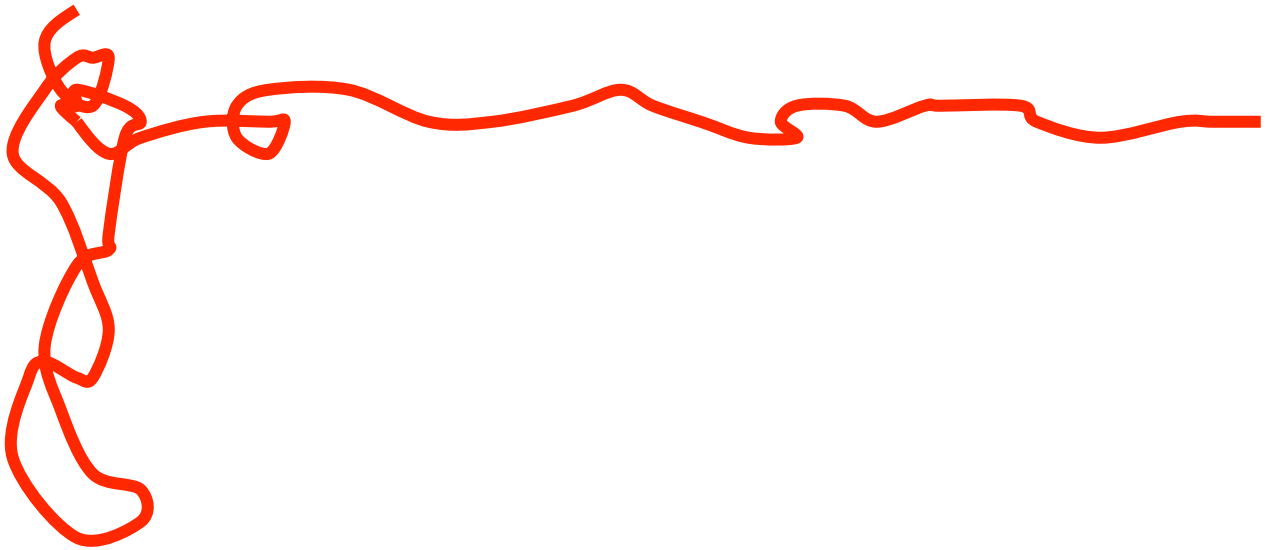


Speech recognition



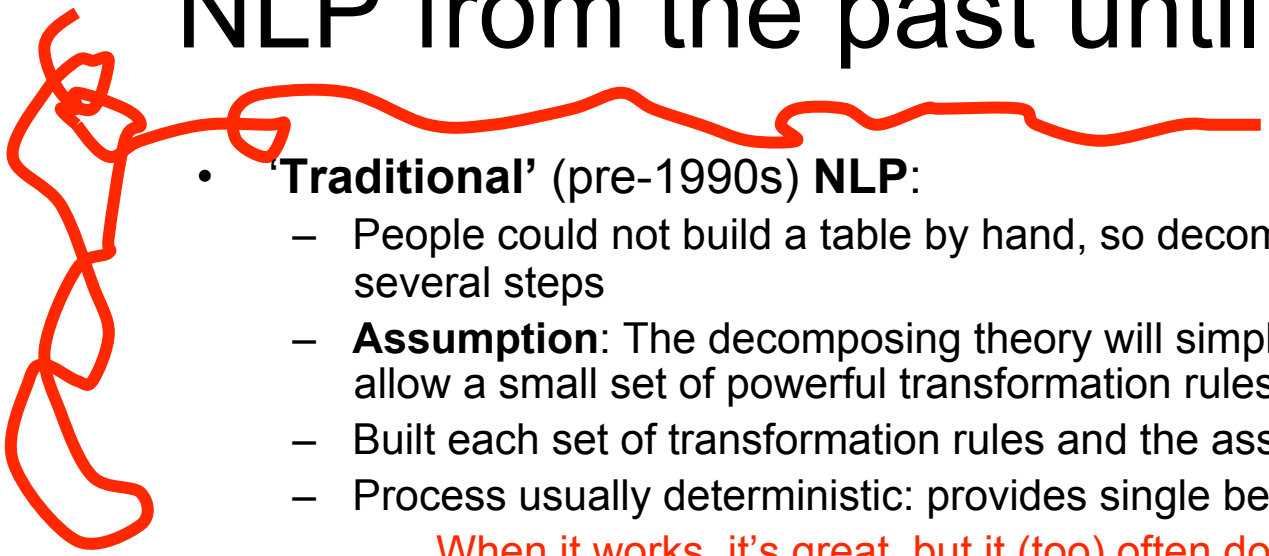
Info Extraction, Text Summ, QA





Where next?

NLP from the past until today



- **'Traditional'** (pre-1990s) **NLP**:
  - People could not build a table by hand, so decomposed the process into several steps
  - **Assumption**: The decomposing theory will simplify the problem enough to allow a small set of powerful transformation rules
  - Built each set of transformation rules and the associated engines by hand
  - Process usually deterministic: provides single best (?) answer, or fails

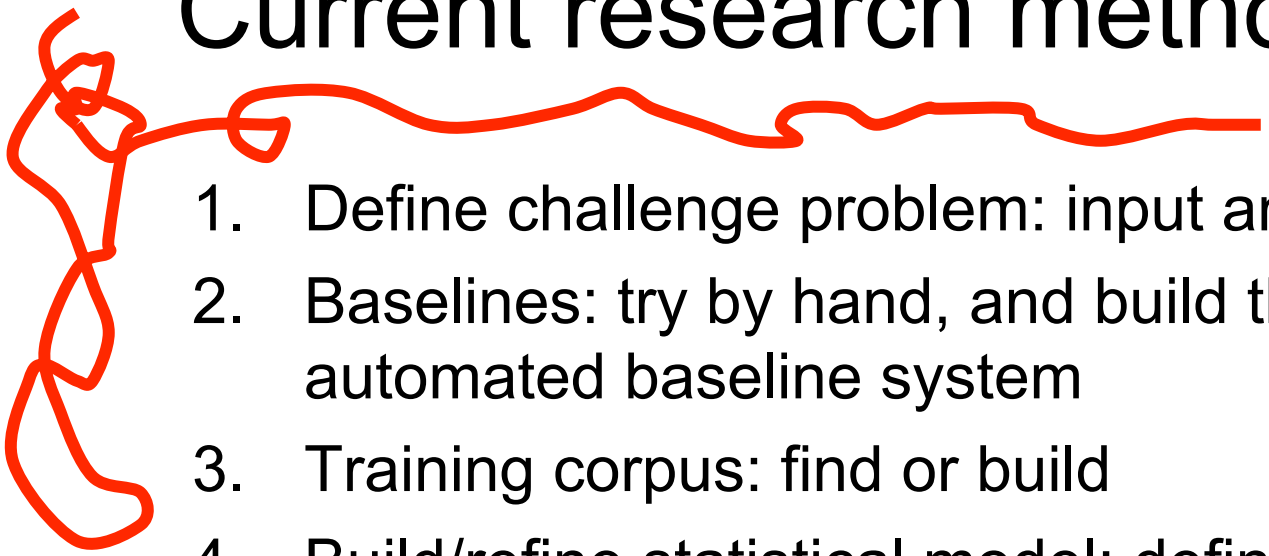
    When it works, it's great, but it (too) often doesn't work

    BUT: building the rules is lots of work!
- **Statistical** (post-1990s) **NLP**:
  - People build tables as probabilized transformations automatically, using machine learning on corpora
  - **Assumption**: The phenomena are too complex anyway, and a machine is better at learning thousands of interdependencies
  - Initially, people tried one-step transformations (e.g., IBM's word-replacement MT); later, started decomposing process as well, with slightly different decomposition choices
  - Process usually provides multiple answers, ranked, and seldom completely fails

    It usually works, but sometimes provides bad results

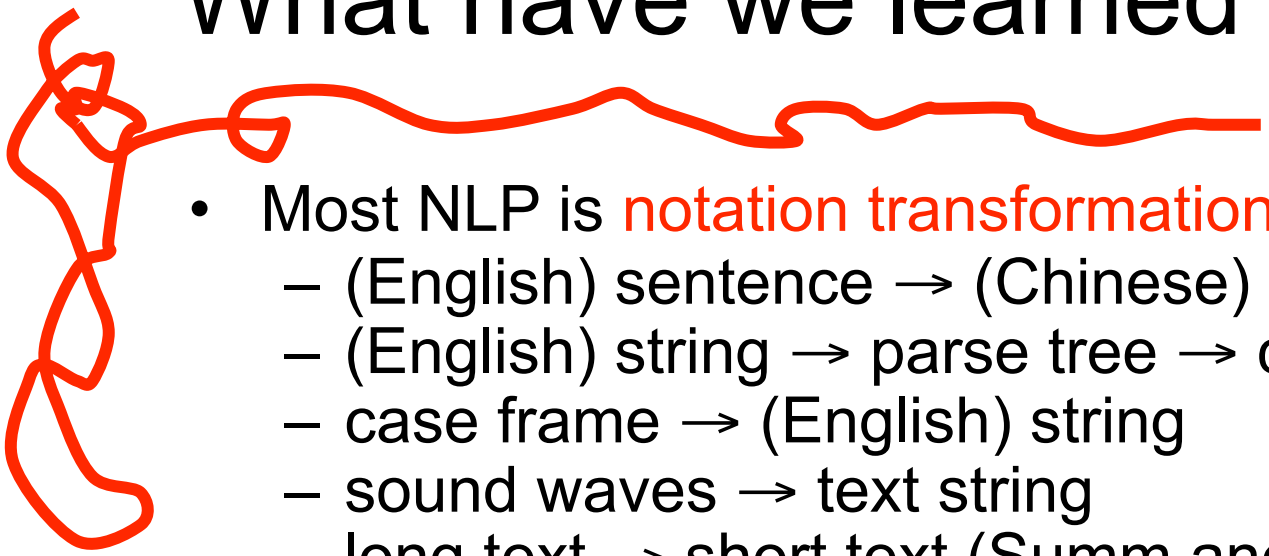
    BUT: building the corpus is lots of work!

Current research methodology



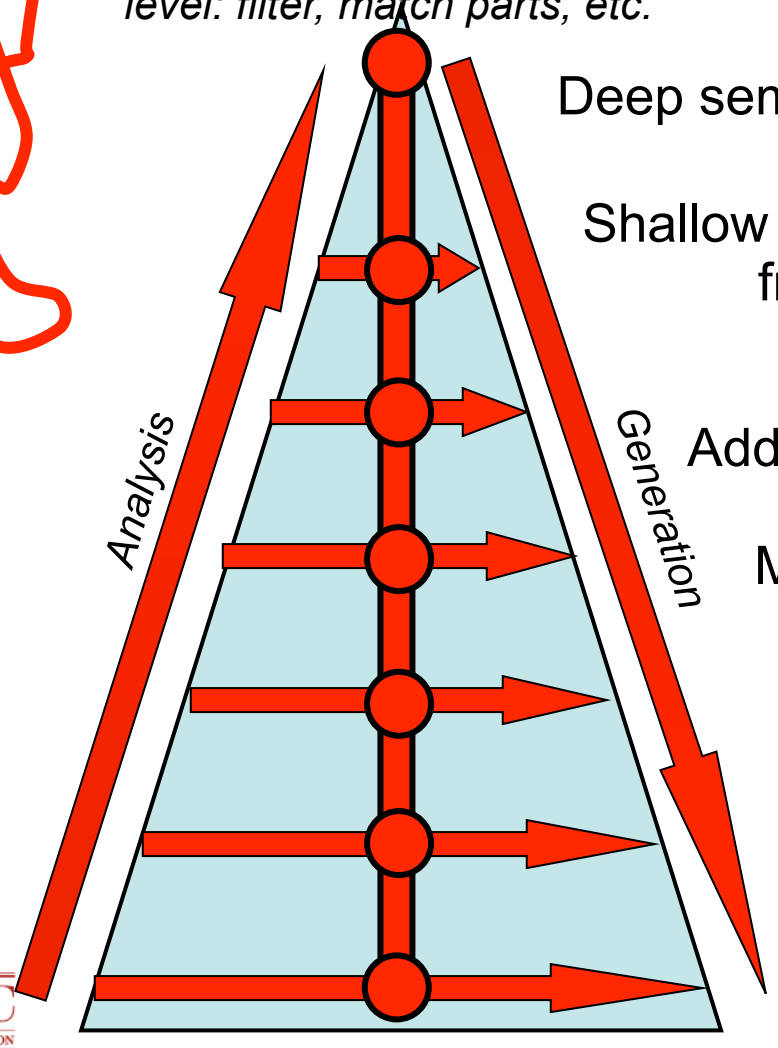
1. Define challenge problem: input and desired output(s)
2. Baselines: try by hand, and build the simplest automated baseline system
3. Training corpus: find or build
4. Build/refine statistical model: define features, define combination function(s)
5. Apply learning algorithms: Naïve Bayes, C4.5, SVM, MaxEnt…; then do bagging, boosting…
6. Evaluate
7. Repeat from step 3 until you beat the best work so far
8. Publish (and maybe put code on your website)

What have we learned about NLP?



- Most NLP is notation transformation:
  - (English) sentence → (Chinese) sentence
  - (English) string → parse tree → case frame
  - case frame → (English) string
  - sound waves → text string
  - long text → short text (Summ and QA)
- …with (often) some information added:
  - POS, syntactic, semantic, and other labels; brackets
  - associated documents
- A little NLP is theorizing:
  - designing the notation model: level and formalism
- Much NLP is engineering:
  - Selecting and tuning learning performance — (rapid) build-evaluate-build cycle

A hierarchy of transformations



Deep semantics: ?

Shallow semantics: frames

Adding more: semantic features

Mid-level changes: syntax

Adding info: POS tags, etc.

Small changes: demorphing, etc.

Direct: simple replacement

- Some transforms are 'deeper' than others
- Each layer of abstraction defines classes/types of behavioral regularity
- These types solve the data sparseness problem

More phenomena of semantics

Somewhat easier

- Bracketing (scope) of predications
- Word sense selection (incl. copula)
- NP structure: genitives, modifiers…
- Concepts: ontology definition
- Concept structure (incl. frames and thematic roles)
- Coreference (entities and events)
- Pronoun classification (ref, bound, event, generic, other)
- Identification of events
- Temporal relations (incl. discourse and aspect)
- Manner relations
- Spatial relations
- Direct quotation and reported speech

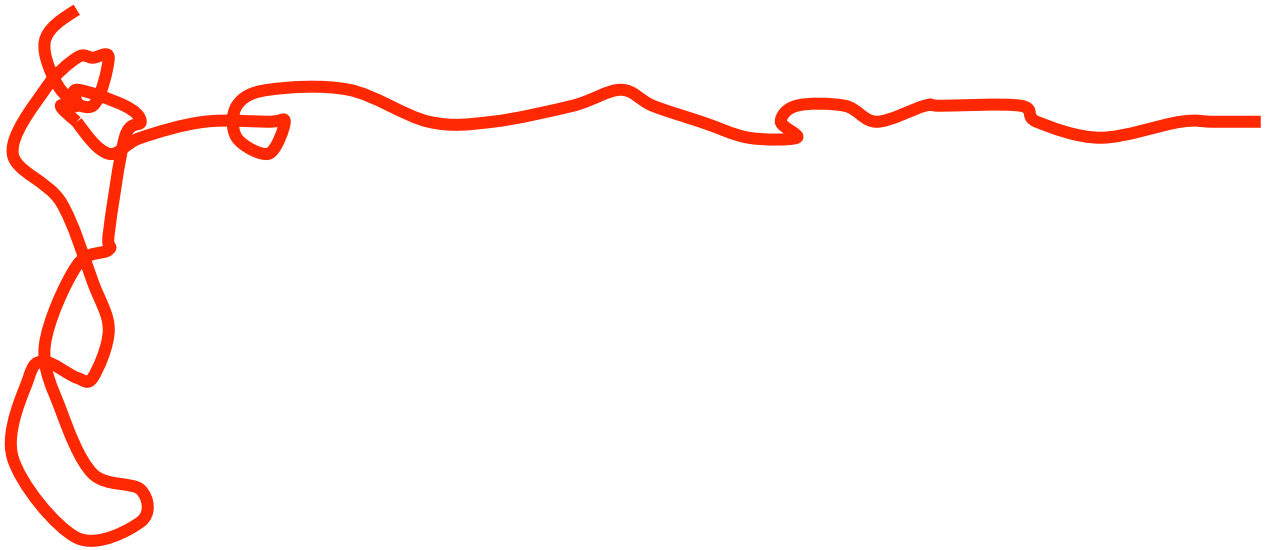
More difficult / 'deeper'

- Quantifier phrases and numerical expressions
- Comparatives
- Coordination
- Information structure (theme/rheme)
- Focus
- Discourse structure
- Other adverbials (epistemic modals, evidentials)
- Identification of propositions (modality)
- Pragmatics/speech acts
- Polarity/negation
- Presuppositions
- Metaphors

Building the layers and their engines

- Some layers are well-understood:
  - Morphology: analyzers essentially 100%
  - POS: taggers at 96%+
  - Syntax: parsers at 90%+
  - Entities: extractors at 65–85%
- Some are only now being explored in NLP:
  - Shallow semantics: no theories, no wide-coverage engines, no large-scale corpora
  - Info structure: probably doable
  - (Co-)reference: engines at 65% and improving…
  - Opinion (judgments and beliefs): simple cases
  - Entailments: starting to learn relevant features…
- Some are too advanced for large-scale robust processing: engines for deep(er) semantics, discourse structure, robust NL generation, dialogue, style…

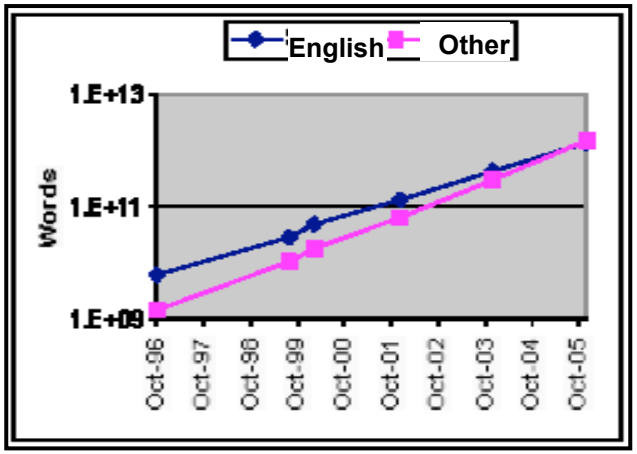
How to get them?



Some applications

The web: A giant opportunity

- Jan 02: over 195 bill words (2 bill pages) on the web (66% English, 12% German, 5% Japanese)
- Need IR, MT, summ, QA
- Need semantics (ontologies)



	Sample (thousands of words)			Exponential Growth Assumption		
Language	Oct-96	Aug-99	Feb-00	Dec-01	Dec-03	Dec-05
English	6,082.09	28,222.10	48,064.10	128,043.57	419,269.14	1,375,098.05
German	228.94	1,994.23	3,333.13	13,435.07	65,161.79	316,727.36
Japanese	228.94	1,994.23	3,333.13	9,375.41	40,070.32	171,600.89
French	223.32	1,529.80	2,732.22	9,375.41	40,070.32	171,600.89
Spanish	104.32	1,125.65	1,894.97	8,786.78	48,968.42	273,542.30
Chinese	123.56	817.27	1,338.35	8,786.78	48,968.42	273,542.30
Korean	123.56	817.27	1,338.35	4,507.93	18,206.81	73,675.11
Italian	123.56	817.27	1,338.35	4,507.93	18,206.81	73,675.11
Portuguese	106.17	589.39	1,161.90	3,455.98	13,438.26	52,350.71
Norwegian	106.50	669.33	947.49	3,109.04	11,474.59	42,425.27
Finnish	20.65	107.26	166.60	480.19	1,628.87	5,534.62
Non-English	1,389.49	10,461.70	17,584.48	65,820.52	306,194.61	1,454,674.58
Non-English%	18.60%	27.04%	26.79%	33.95%	42.21%	51.41%

(from Grefenstette 99, with additions by Oard and Hovy)

USC INFORMATION SCIENCES INSTITUTE

ISI Information Sciences Institute

The Semantic Web: dream

- Strong vision: each webpage (text, picture, graph, etc.) supported by semantic (Interlingual) description; search engines use this; presentation engines translate into user's language
- Problems:
  - Automated description creation from text: requires semantic analysis!
  - Automated description creation from other media: who knows?
  - Standardized Interlingua termset / ontology: how many terms? Who will make them?
  - Automatic presentation generators: fluent multi-sentence multi-lingual generation is a still a dream

…so is the Semantic Web just a dream?

The Semantic Web: reality

- Weak vision: each webpage contains (semantic) annotations; search and display engines use them
- Problems:
  - How to find critical indexing terms? (Cranfield experiments!)
  - What to do with non-text media?
  - Which terms? Which terminology standard?
  - How to display results?

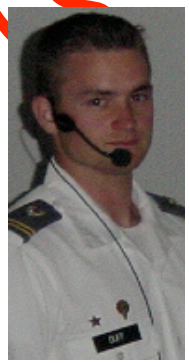
Can do better than Google; CLIR in TREC

Captions; graph interpretation

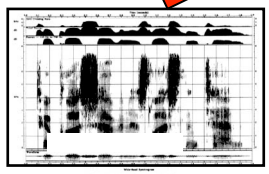
Use WordNet or others; create term converters

Link to MT engines

Deep focus: Complex agents with NL



from sounds…



…to words…

"will you move the clinic there?"

…to meanings…

speech-act <a123>
action *info-request*
addressee *doctor*
content <v27>
type *question*
q-slot *polarity*
prop <p99>
type action
event *move*
agent *doctor*
theme *clinic*
destination *there*
time *future*

…integrating with mental state…

Dialogue & thinking:
Updating beliefs, goals, emotions, intentions,…

…forming communicative goals…

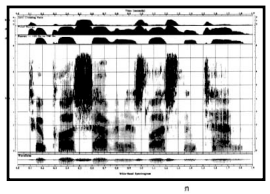
speech-act <a124>
action *answer*
addressee *captain*
***question* <v27>**
answer <v34>
type *assert*
content <p102>
type *move*
agent *doctor*
theme *clinic*
destination *there*
time *future*
polarity *positive*

from meanings…

…to words…

"yes"

…to sounds



Shallow focus: Many applications

- Handheld tourist assistant
  - speech+translation+multimedia
  - travel info, maps…
- Online business news
  - new news (novelty)+headline summarization
- Info gathering, for report writer & education
  - complex QA+summarization
- Semantic web usage, for everyone
  - parsing+keyword MT+multi-ling generation
  - Google++…

Where next?

☺ By creating smaller transformation (sub)steps, we can learn better, and branch out to more apps:

- define a new cutlevel (with notation!),
- list many X→Y pairs as training data
- learn the transformation rules.

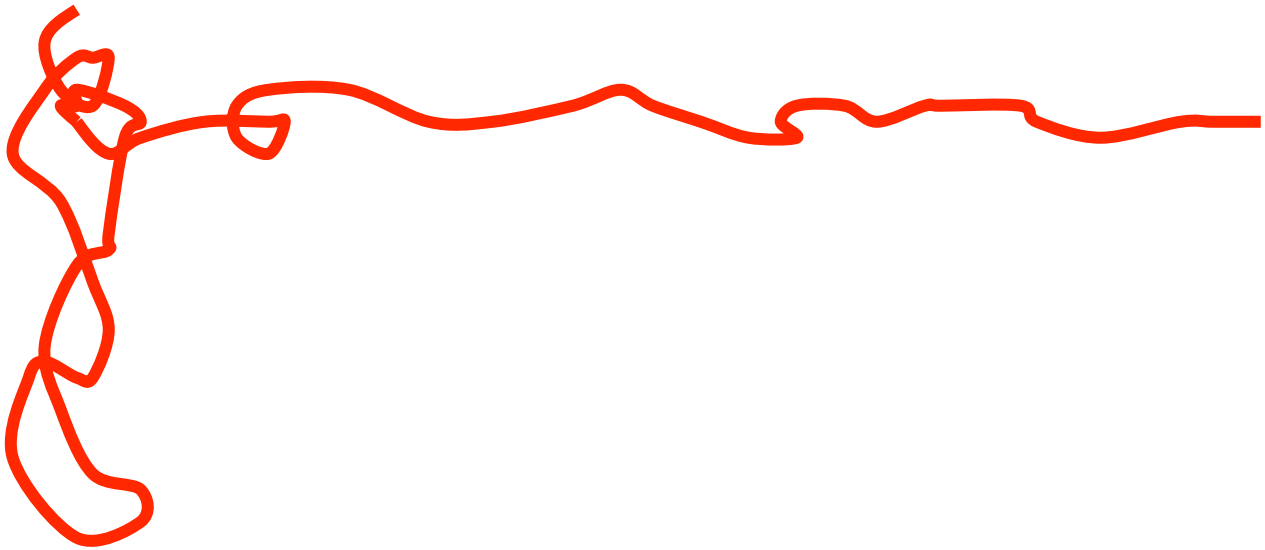
add info that isn't in the text

use EM, ME, etc. to learn best 'rules'

☹ Major bottlenecks:

- Diffuse phenomena require very large training sets: algorithms and speed issues
- Shallow semantics
- Discourse and dialogue
- Pragmatics and hearer/user models
- Information Theory

cutlevels



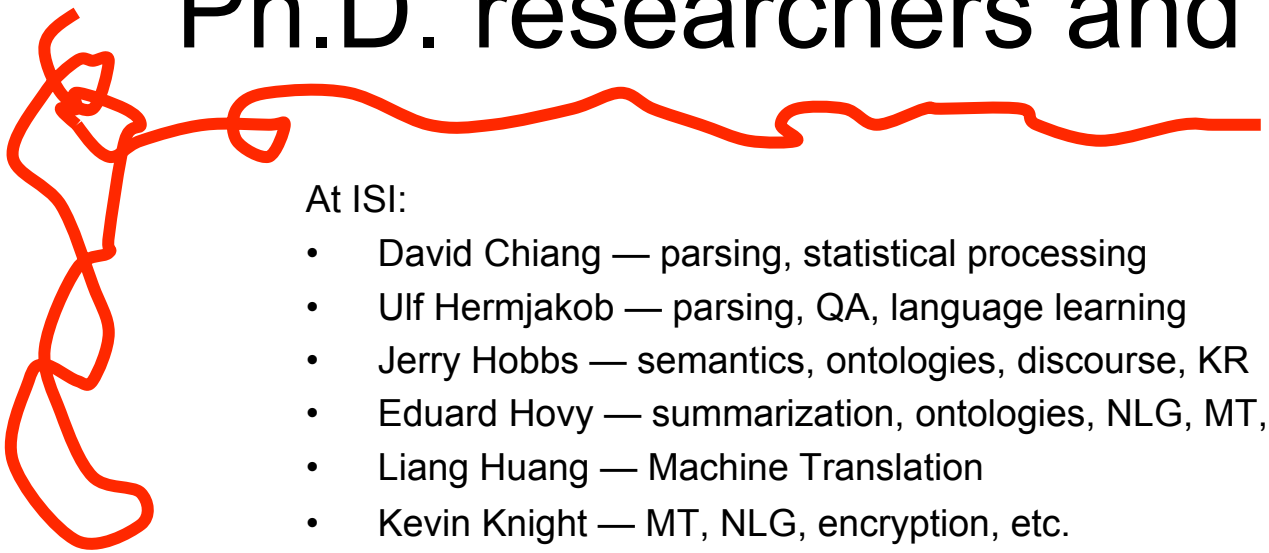
NLP at USC: ISI and ICT

USC INFORMATION SCIENCES INSTITUTE

ISI Information Sciences Institute


Information Sciences Institute
USC
INFORMATION
SCIENCES
INSTITUTE
UNIVERSITY OF SOUTHERN CALIFORNIA
School of Engineering

Ph.D. researchers and topics



At ISI:

- David Chiang — parsing, statistical processing
- Ulf Hermjakob — parsing, QA, language learning
- Jerry Hobbs — semantics, ontologies, discourse, KR
- Eduard Hovy — summarization, ontologies, NLG, MT, etc.
- Liang Huang — Machine Translation
- Kevin Knight — MT, NLG, encryption, etc.
- Zornitsa Kozareva — Information Extraction, text mining
- Daniel Marcu — QA, summarization, discourse, etc.
- (Patrick Pantel — clustering, ontologies, learning by reading)

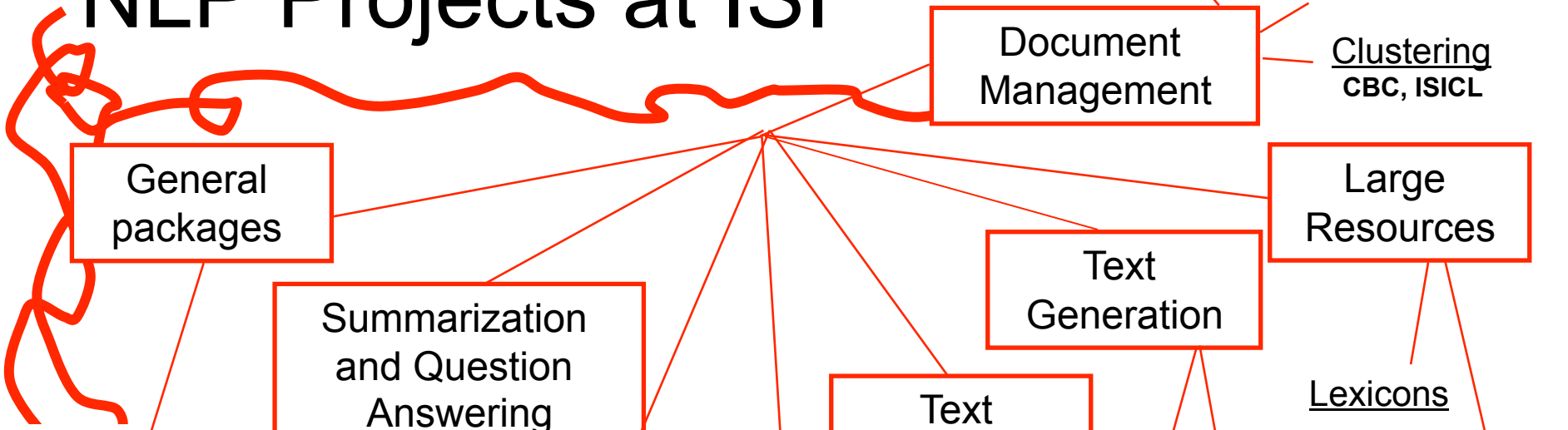
At ICT:

- David DeVault — NL generation
- Andrew Gordon — cognitive science and language
- Anton Leuski — IR
- Kenji Sagae — parsing
- Bill Swartout — NLG
- David Traum — dialogue

At USC/EE:

- Shri Narayanan — speech recognition

NLP Projects at ISI



Social Network Analysis
Email analysis

Web Access / IR
MuST / C*ST*RD

Document Management

Clustering
CBC, ISICL

General packages

Large Resources

Summarization and Question Answering

Text Generation

Text Analysis

Lexicons

Information Extraction

Ontologies
OntoNotes (semantic corpus)
Omega (for MT, summarization)
DINO (for multi-database access)
CORCO (semi-auto construction)

Machine Translation

EM: **YASMET**
MT: **GIZA**
FSM: **CARMEL**
Name transliteration
Clustering: **ISICL**
Single-doc: **SUMMARIST (English, Spanish, Indonesian, German)**
Multi-doc: **NeATS, AGILE compaction, GOSP (headlines)**
Evaluation: **SEE, ROUGE, BE Breaker**

Med. informatics
Psyop/SOCOM
Learning by Reading
eRulemaking

TEXTMAP (English)
WEBCLOPEDIA (English, Korean, Chinese)

AGILE (Arabic, Chinese)
REWRITE (Chinese, Arabic, Tetun)
TRANSONIC (speech translation)
ADGEN
GAZELLE (Japanese, Spanish, Arabic)
QuTE (Indonesian)

Text Planning
Sent. Planning
ICT agent-based NLP
HealthDoc

Sent. Realization
NITROGEN, HALOGEN
PENMAN

Discourse Parsing
DMT (English, Japanese)

Sentence Parsing, Grammar Learning
CONTEX (English, Japanese, Korean, Chinese)
Parser and grammar learning