



CS114 Introduction to Computational Linguistics

Marie Meteer

January 13, 2014

Brandeis University

Additional slides courtesy of Jurafsky & Martin, James Pustejovsky and , Ray Mooney

Course Details

- Web site: www.cs.brandeis.edu/~cs114
- My email: mmeteer@cs.brandeis.edu
- Course TA: John Vogel, locuscosecant@gmail.com
- Textbook:
 - *Speech and Language Processing*, Jurafsky and Martin
- Programming assignments in Python
 - Text: *NLP with Python*, Bird, Klein, and Loper
 - Python tutorial: TBD

Natural Language Processing

- We're going to study what goes into getting computers to perform useful and interesting tasks involving human languages.
- We are also concerned with the insights that such computational work gives us into human processing of language.

Why Should You Care?

Two trends

1. An enormous amount of knowledge is now available in machine readable form as natural language text
2. Conversational agents are becoming an important form of human-computer communication
3. Much of human-human communication is now mediated by computers

Commercial World

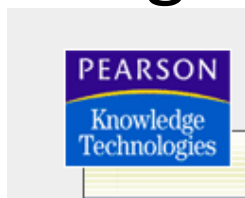
- A lot of opportunities in CL ...



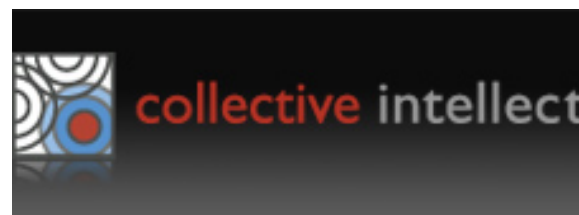
Commercial World

- Lot's of exciting stuff going on...

Microsoft[®]



J.D. POWER
AND ASSOCIATES[®]



Google Translate

الجزيرة نت

http://www.aljazeera.net/NR/exeres/8FD54E7F-56C5-49A0-B60A-89A67426F3B3.htm al jazeera

Speech and ...of Contents Book Schedule University o...uter Science James Marti... Home Page The Daily Camera The New Yor... Multimedia

نحضر لك الأخبار الساخنة أينما تكون
اشترك الآن

ملخص الخبر

انتقل الأمين العام للأمم المتحدة الذي يزور لبنان لقرعة المراقبة الدولية ببغداد الناقورة الجنوبية حيث وضع إكليلًا من الزهور على نصب لاربعه مراقبين قتلوا خلال الحرب الأخيرة. وفيما ينتظر تفقد آنان للقرى المدمرة تنظم إيطاليا اليوم احتفالا لتوديع قوتها المرسله إلى لبنان.

أنا في الناقورة لتفقد الخط الأزرق والمناطق للمدرة

الثلاثاء 5/8/1427 هـ - الموافق 29/8/2006 م (آخر تحديث) الساعة 17:44 (مكة المكرمة)، 14:44 (غرينتش)

مكة إسرائيلية تمدد امتثال وزير للالاية الفلسطينية و13 نائبا من حماس

استشهاد فلسطينيين وإصابة تسعة في غارات بالضفة والقطاع

أصيب تسعة فلسطينيين بينهم مدنيون في غارة جوية إسرائيلية على حي الشجاعية في قطاع غزة. يأتي ذلك مباشرة بعد استشهاد قنايين بارزين من كتائب شهداء الأقصى في عملية لقوات الاحتلال الإسرائيلي نفذها سلاح الجو وقوات المشاة في مخيم بلاطة بالضفة الغربية.

البشير يلتقي فريزر ومجلس الأمن لن يفرض قوات بدارفور

من المقرر أن يلتقي الرئيس السوداني عمر البشير جيندائي فريزر مساعدة وزيرة الخارجية الأميركية التي تحاول في الخرطوم إقناع المسؤولين السودانيين بنشر قوة أممية بدارفور. من جانبه قال السفير الأميركي في الأمم المتحدة إنه لا نية لمجلس الأمن بفرض قوات في الإقليم.

رمسفيلد وتشيني يصران على إبقاء القوات الأميركية بالعراق

دعا وزير الدفاع الأميركي دونالد رمسفيلد الأميركيين إلى التحلي بالصبر بخصوص العراق. وانتهد ديك تشيني نائب الرئيس دعوات الديمقراطيين لسحب القوات الأميركية من العراق للربط بين الانسحاب المبكر واحتمال وقوع هجمات داخل الولايات المتحدة.

مقتل مدنيين وإصابة ضابط بهجوم انتحاري في أفغانستان

أعلنت القوة الدولية للمساعدة على إرساء الأمن (إيساف) مقتل مدنيين وإصابة ضابط أفغاني في هجوم استهدف قافلة لقوات الأطلسي بجنوب أفغانستان. وفي العاصمة كابل انفجرت قنبلة يدوية الصنع لدى مرور بعثة فرنسية من أجل تقديم المساعدة الإنسانية لضحايا.

الجزيرة نت
ALJAZEERA.NET

الأخبار الفضائية المعرفة الأعمال

روابط أخرى

بحث بحث تفصيلي خدمات الموقع

انذهب

غزو غزة بالأيام

هذه الصفحة برعاية

المتفوقون الحقيقيون لا يلحقون بالأخريين بل يقودون الطريق

القطرية QATAR AIRWAYS

اسم البرنامج: بلا حدود

عنوان الحلقة: يوسف دنا.. اتهامات بدعم الإرهاب

الأربعاء/ مباشر

19:05 غرينتش 22:05 مكة

frontiers@aljazeera.net

التفصيلية الاقتصادية الرياضية

Web Q/A



The screenshot shows a search interface with the following elements:

- Search Bar:** Contains the text "what's the population of boulder" and a magnifying glass icon.
- Navigation:** A horizontal bar with tabs for "Web", "Images", "News", "Maps", "QnA ^{Beta}", and "More ▾".
- Results:** A single result is displayed: **Boulder, Colorado Population, total: 92,196**. Below the title, it says "2004 estimate · US Census Bureau". To the right of the title is a link "Is this useful?".
- Page Info:** Below the navigation bar, it says "Page 1 of 112,364 results · [Options](#)".

Weblog Analytics

- Data-mining of Weblogs, discussion forums, message boards, user groups, and other forms of user generated media
 - Product marketing information
 - Political opinion tracking
 - Social network analysis
 - Buzz analysis (what's hot, what topics are people talking about right now).

What makes Language Hard?

- The little words
 - I read a book in Chinese
 - I read a book on Chinese.
- What we don't say
 - I read a book on the subway.
 - I read a book on NY.
 - John started a new book.
 - JK Rowling started a new book.
- How we say it
 - Different meaning
 - The white house is on the corner.
 - The White House is on the fence.
 - Focusing attention on corrections
 - We will send the response to Donald Metzger
 - No that's Ronald Metzger vs. No, that's Donald Fetzger
 - Emphasizing important information
 - I need to be in Boston no later than 9 am.

What Makes Speech Hard

- Confusability
 - 4 2 8 vs. 4 to 8
 - let us pray vs. lettuce spray
 - It is easy to recognize speech vs. It is easy to wreck a nice beach
 - A new display vs. a nudist play
 - Alienate peter vs. a lion ate peter
- Coarticulation
 - Going to vs gonna
 - Dish soap

More ...

- Garden Path Constructions
 - *The horse raced past the garden fell.*
 - *The old man the boat.*
- Ambiguity
 - Semantics
 - *I saw the boy with a {telescope / tattoo}*
 - Pragmatics
 - *((old men) and women) as opposed to (old men and women) in “Old men and women were taken to safe location”, since women- both and young and old- were very likely taken to safe locations*
 - Discourse:
 - *No smoking areas allow hookahs inside, except the one in Hotel Grand.*
 - *No smoking areas allow hookahs inside, but not cigars.*

Major Topics

1. Words
2. Syntax
3. Meaning
4. Discourse
5. Context

5. Applications exploiting each

- Question answering
- Conversational agents
- Information extraction
- Summarization
- Machine translation

History: foundational insights 1940s-1950s

- Automaton:
 - Turing 1936
 - McCulloch-Pitts neuron (1943)
 - <http://diwww.epfl.ch/mantra/tutorial/english/mcpits/html/>
 - Kleene (1951/1956)
 - Shannon (1948) link between automata and Markov models
 - Chomsky (1956)/Backus (1959)/Naur(1960): CFG
- Probabilistic/Information-theoretic models
 - Shannon (1948)
 - Bell Labs speech recognition (1952)

History: the two camps: 1957-1970

- Symbolic
 - Zellig Harris 1958 TDAP first parser?
 - Transformations and Discourse Analysis Project
 - Cascade of finite-state transducers
 - Chomsky, Syntactic Structures, 1957
 - Newell and Simon: Logic Theorist, General Problem Solver
 - 1956 computer program that could prove mathematical theorems
 - Dartmouth Summer Research Project on Artificial Intelligence 1958
 - McCarthy, Minsky, Shannon, Rochester
 - “The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.”
- Statistical
 - Bledsoe and Browning (1959): Bayesian OCR
 - Mosteller and Wallace (1964): Bayesian authorship attribution
 - Denes (1959): ASR combining grammar and acoustic probability

Four paradigms: 1970-1983

- Stochastic
 - Hidden Markov Model 1972
 - Independent application of Baker (CMU) and Jelinek/Bahl/Mercer lab (IBM) following work of Baum and colleagues at IDA
- Logic-based
 - Colmerauer (1970,1975) Q-systems
 - Definite Clause Grammars (Pereira and Warren 1980)
 - Kay (1979) functional grammar, Bresnan and Kaplan (1982) unification
- Natural language understanding
 - Winograd (1972) Shrdlu
 - Schank and Abelson (1977) scripts, story understanding
 - Influence of case-role work of Fillmore (1968) via Simmons (1973), Schank.
- Discourse Modeling
 - Grosz and colleagues: discourse structure and focus
 - Perrault and Allen (1980) BDI model

Empiricism and Finite State Revival: 1983-1993

- Finite State Models
 - Kaplan and Kay (1981): Phonology/Morphology
 - Church (1980): Syntax
- Return of Probabilistic Models:
 - Corpora created for language tasks
 - Early statistical versions of NLP applications (parsing, tagging, machine translation)
 - Increased focus on methodological rigor:
 - Can't test your hypothesis on the data you used to build it!
 - Training sets and test sets

The field comes together: 1994-2011

- NLP has borrowed statistical modeling from speech recognition, is now standard:
 - ACL conference:
 - 1990: 39 articles 1 statistical
 - 2003 62 articles 48 statistical
 - Machine learning techniques key
- NLP has borrowed focus on web and search and “bag of words models” from information retrieval
- Unified field:
 - NLP, MT, ASR, TTS, Dialog, IR

Ambiguity

- Ambiguity is a fundamental feature of natural communicative systems.
- Identifying and resolving ambiguity is a critical goal in understanding language.
- Most current CL work involves ambiguity in some way.

Ambiguity

- Find at least 5 meanings of this sentence:
 - I made her duck.

Ambiguity

- Find at least 5 meanings of this sentence:
 - I made her duck.
- I cooked waterfowl for her benefit (to eat)
- I cooked waterfowl belonging to her
- I created the (plaster?) duck she owns
- I caused her to quickly lower her head or body
- I waved my magic wand and turned her into undifferentiated waterfowl

Ambiguity is Pervasive

- I caused her to quickly lower her head or body
 - **Lexical category**: “duck” can be a N or V
- I cooked waterfowl belonging to her.
 - **Lexical category**: “her” can be a possessive (“of her”) or dative (“for her”) pronoun
- I made the (plaster) duck statue she owns
 - **Lexical Semantics**: “make” can mean “create” or “cook”

Ambiguity is Pervasive

- **Grammar: Make can be:**
 - **Transitive: (verb has a noun direct object)**
 - I cooked [waterfowl belonging to her]
 - **Ditransitive: (verb has 2 noun objects)**
 - I made [her] (into) [undifferentiated waterfowl]
 - **Action-transitive (verb has a direct object and another verb)**
 - I caused [her] [to move her body]

Ambiguity is Pervasive

- **Phonetics!**
 - I mate or duck
 - I'm eight or duck
 - Eye maid; her duck
 - Aye mate, her duck
 - I maid her duck
 - I'm aid her duck
 - I mate her duck
 - I'm ate her duck
 - I'm ate or duck
 - I mate or duck

Dealing with Ambiguity

- Four possible approaches:
 1. Tightly coupled interaction among processing levels; knowledge from other levels can help decide among choices at ambiguous levels (e.g. “blackboard” architecture).
 2. Pipeline processing that ignores ambiguity as it occurs and hopes that other levels can eliminate incorrect structures.

Dealing with Ambiguity

3. Probabilistic approaches based on making the most likely choices
4. Don't do anything, maybe it won't matter
 1. *We'll leave when the duck is ready to eat.*
 2. *The duck is ready to eat now.*
 - Does the “duck” ambiguity matter with respect to whether we can leave?

Categories of Knowledge

- Phonology
- Morphology
- Syntax
- Semantics
- Pragmatics
- Discourse

Each kind of knowledge has associated with it an encapsulated set of processes that make use of it.

Interfaces are defined that allow the various levels to communicate.

This usually leads to a pipeline architecture.

Word Segmentation

- Breaking a string of characters (graphemes) into a sequence of words.
- In some written languages (e.g. Chinese) words are not separated by spaces.
- Even in English, characters other than white-space can be used to separate words [e.g. , ; . - : ()]
- Examples from English URLs:
 - jumptheshark.com ⇒ jump the shark .com
 - myspace.com/pluckerswingbar
 - ⇒ myspace .com pluckers wing bar
 - ⇒ myspace .com plucker swing bar

Morphological Analysis

- **Morphology** is the field of linguistics that studies the internal structure of words. (Wikipedia)
- A **morpheme** is the smallest linguistic unit that has semantic meaning (Wikipedia)
 - e.g. “carry”, “pre”, “ed”, “ly”, “s”
- Morphological analysis is the task of segmenting a word into its morphemes:
 - carried \Rightarrow carry + ed (past tense)
 - independently \Rightarrow in + (depend + ent) + ly
 - Googlers \Rightarrow (Google + er) + s (plural)
 - unlockable \Rightarrow un + (lock + able) ?
 \Rightarrow (un + lock) + able ?

Part Of Speech (POS) Tagging

- Annotate each word in a sentence with a part-of-speech.

I ate the spaghetti with meatballs.

Pro V Det N Prep N

John saw the saw and decided to take it to the table.

PN V Det N Con V Part V Pro Prep Det N

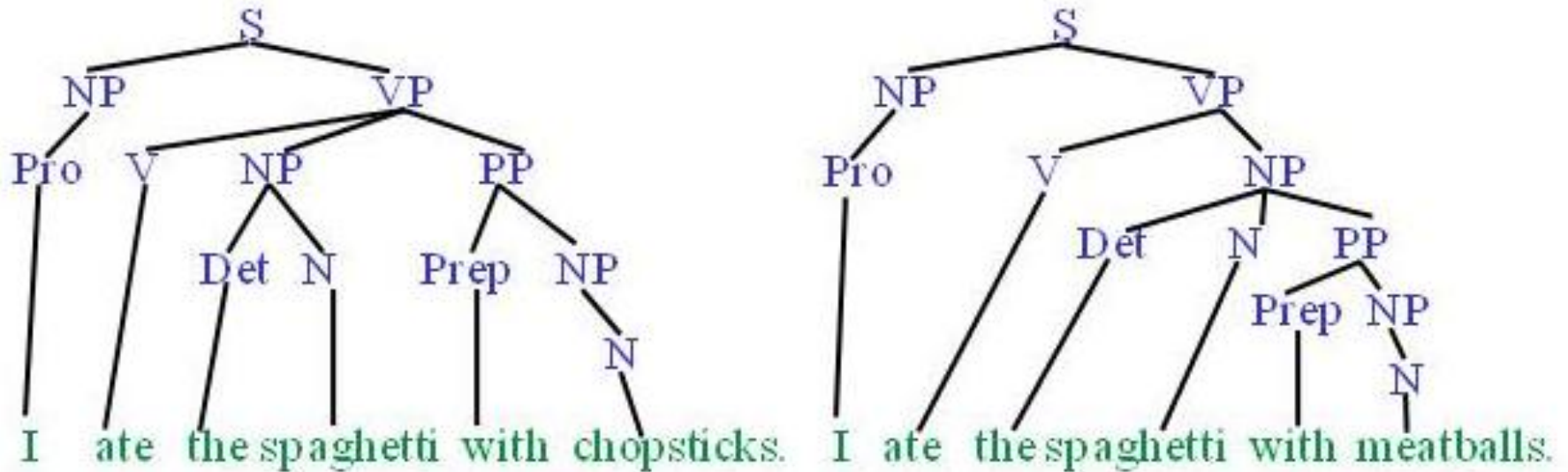
- Useful for subsequent syntactic parsing and word sense disambiguation.

Phrase Chunking

- Find all non-recursive noun phrases (NPs) and verb phrases (VPs) in a sentence.
 - [NP I] [VP ate] [NP the spaghetti] [PP with] [NP meatballs].
 - [NP He] [VP reckons] [NP the current account deficit] [VP will narrow] [PP to] [NP only \$1.8 billion] [PP in] [NP September]

Syntactic Parsing

- Produce the correct syntactic parse tree for a sentence.



Word Sense Disambiguation (WSD)

- Words in natural language usually have a fair number of different possible meanings.
 - Ellen has a strong interest in computational linguistics.
 - Ellen pays a large amount of interest on her credit card.
- For many tasks (question answering, translation), the proper sense of each ambiguous word in a sentence must be determined.

Semantic Role Labeling (SRL)

- For each clause, determine the semantic role played by each noun phrase that is a participant in the verbal event.

agent patient source destination instrument

– John drove Mary from Austin to Dallas in his Toyota Prius.

– The hammer broke the window.

- Also referred to a “case role analysis,” “thematic analysis,” and “shallow semantic parsing”



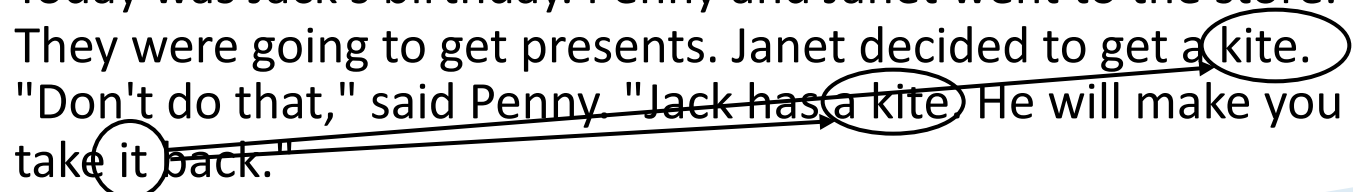
Semantic Parsing

- A *semantic parser* maps a natural-language sentence to a complete, detailed semantic representation (*logical form*).
- For many applications, the desired output is immediately executable by another program.
- Example: Mapping an English database query to Prolog:

How many cities are there in the US?

```
answer(A, count(B, (city(B), loc(B, C),  
                    const(C, countryid(USA))),  
A))
```

Anaphora Resolution/ Co-Reference

- Determine which phrases in a document refer to the same underlying entity.
 - John put the carrot on the plate and ate it.
 - Bush started the war in Iraq. But the president needed the consent of Congress.
- Some cases require difficult reasoning.
 - Today was Jack's birthday. Penny and Janet went to the store. They were going to get presents. Janet decided to get a kite. "Don't do that," said Penny. "Jack has a kite. He will make you take it back."

Information Extraction (IE)

- Identify phrases in language that refer to specific types of entities and relations in text.
- Named entity recognition is task of identifying names of people, places, organizations, etc. in text.

people organizations places

– Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

- Relation extraction identifies specific relations between entities.

– Michael Dell is the CEO of Dell Computer Corporation and lives in Austin Texas.

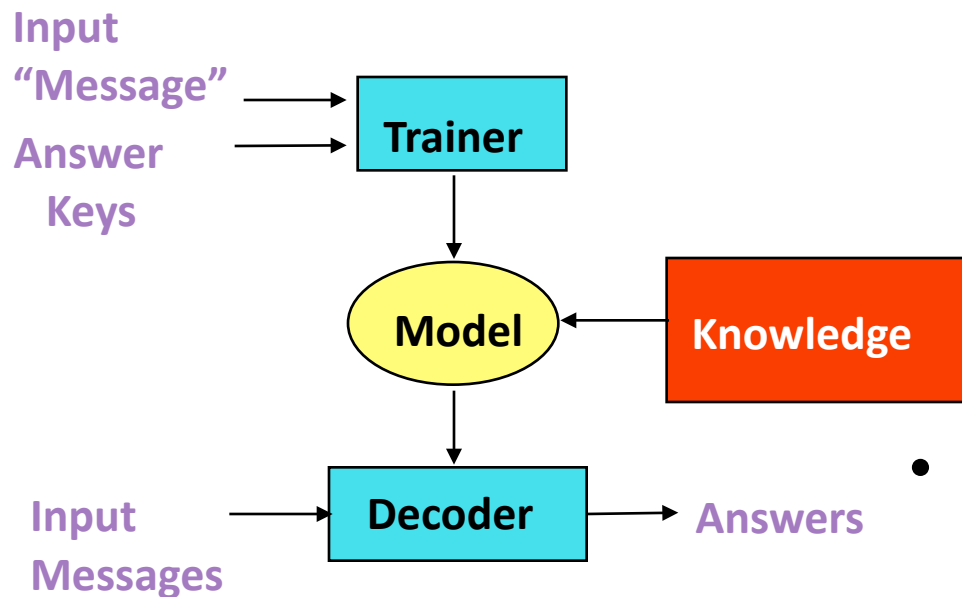
Manual Knowledge Acquisition

- Traditional, “rationalist,” approaches to language processing require human specialists to specify and formalize the required knowledge.
- Manual knowledge engineering, is difficult, time-consuming, and error prone.
- “Rules” in language have numerous exceptions and irregularities.
 - “All grammars leak.”: Edward Sapir (1921)
- Manually developed systems were expensive to develop and their abilities were limited and “brittle” (not robust).

Automatic Learning Approach

- Use machine learning methods to automatically acquire the required knowledge from appropriately annotated text corpora.
- Various referred to as the “corpus based,” “statistical,” or “empirical” approach.
- Statistical learning methods were first applied to speech recognition in the late 1970’s and became the dominant approach in the 1980’s.
- During the 1990’s, the statistical training approach expanded and came to dominate almost all areas of NLP.

Speech and NL Paradigm



- **Requirements:**

- Annotation of messages with keys
- Linguistic and Domain Knowledge
- Statistical Model
- Training Algorithm
- Decoding Algorithm

- **Benefits:**

- Statistical model can combine multiple kinds of information
- Degrades “softly”, finding the most likely answer
- Learns what information is important to make a decision

Supervised Learning for Language Technologies

Technology	Input	Answers
Speech Recognition	Audio	Transcription
Optical Character Recognition	Image	Characters
Topic classification	Document	Topic labels
Information retrieval	Query	Document
Named entity extraction	Text or speech	Names and categories

Advantages of the Learning Approach

- Large amounts of electronic text are now available.
- Annotating corpora is easier and requires less expertise than manual knowledge engineering.
- Learning algorithms have progressed to be able to handle large amounts of data and produce accurate probabilistic knowledge.
- The probabilistic knowledge acquired allows robust processing that handles linguistic regularities as well as exceptions.

The Importance of Probability

- Unlikely interpretations of words can combine to generate spurious ambiguity:
 - “The a are of l” is a valid English noun phrase (Abney, 1996)
 - “a” is an adjective for the letter A
 - “are” is a noun for an area of land (as in hectare)
 - “l” is a noun for the letter l
 - “Time flies like an arrow” has 4 parses, including those meaning:
 - Insects of a variety called “time flies” are fond of a particular arrow.
 - A command to record insects’ speed in the manner that an arrow would.
- Some combinations of words are more likely than others:
 - “vice president Gore” vs. “dice precedent core”
- Statistical methods allow computing the most likely interpretation by combining probabilistic evidence from a variety of uncertain knowledge sources.