Speech and Language Processing

Lecture 1

Chapter 1 of SLP

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

- Lot’s of exciting stuff going on…

Powerset

PEARSON
Knowledge Technologies

Google™

YAHOO!®

Nielsen
BuzzMetrics

cymfony
harnessing influence 2.0™

collective intellect

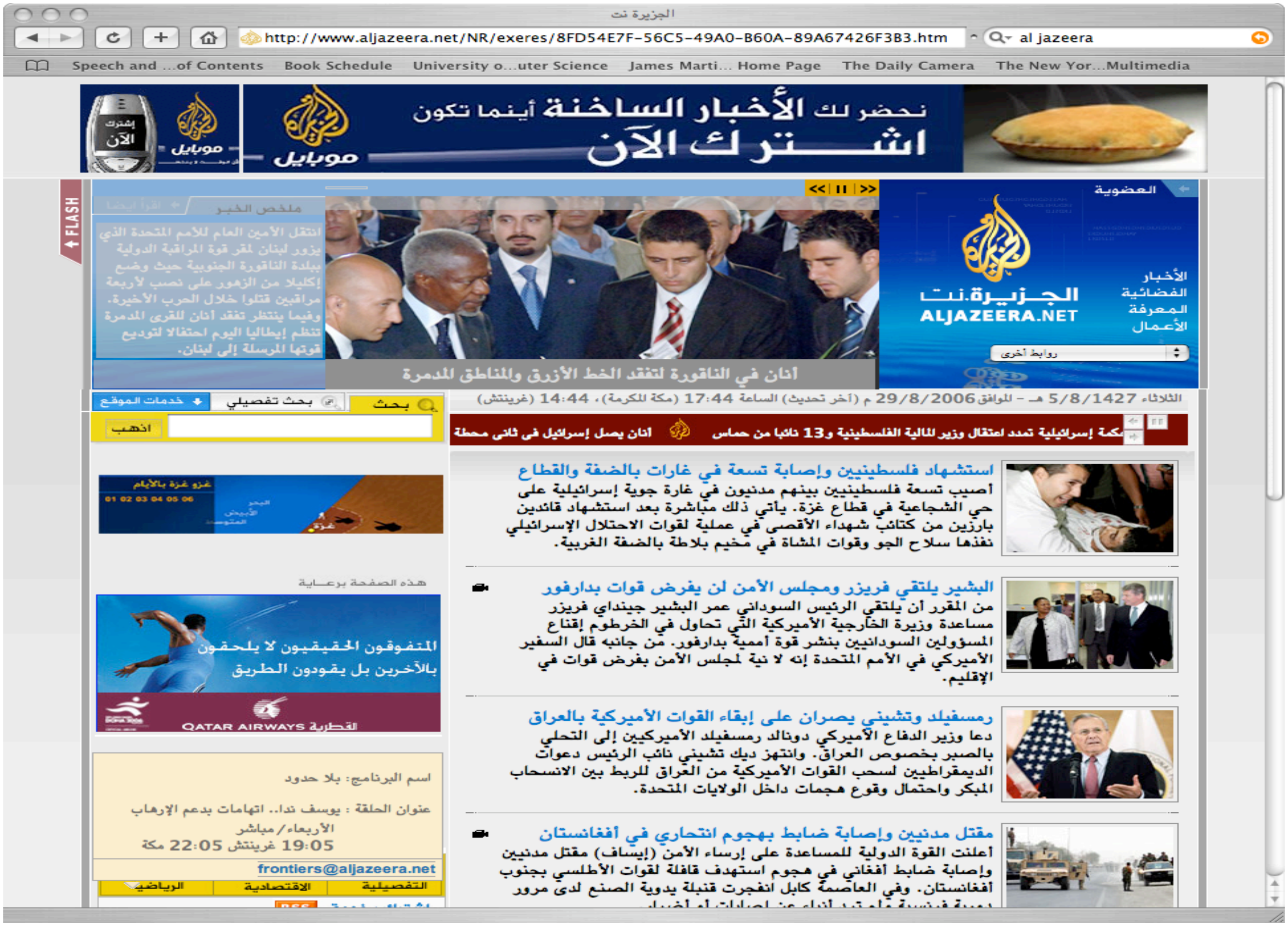
BETA
hakia®
search for meaning

umbria

Commercial World

- Lot’s of exciting stuff going on…

Google Translate



انتقل الأمين العام للأمم المتحدة الذي يزور لبنان لمقر قوة المراقبة الدولية ببلدة الناقورة الجنوبية حيث وضع إكليلا من الزهور على نصب لأربعة مراقبين قتلوا خلال الحرب الأخيرة. وفيما ينتظر تفقد أنان للقرى المدمرة تنظم إيطاليا اليوم احتفالا لتوديع قوتها المرسلة إلى لبنان.

أنان في الناقورة لتفقد الخط الأزرق والمناطق المدمرة

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

كمة إسرائيلية تمدد اعتقال وزير المالية الفلسطينية و13 نائبا من حماس

أنان يصل إسرائيل في ثاني محطة

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 قائدين 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 جينداي 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. وانتهز 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 دورية فرنسية ولم ترد أنباء عن إصابات أو أضرار.

اسم البرنامج: بلا حدود

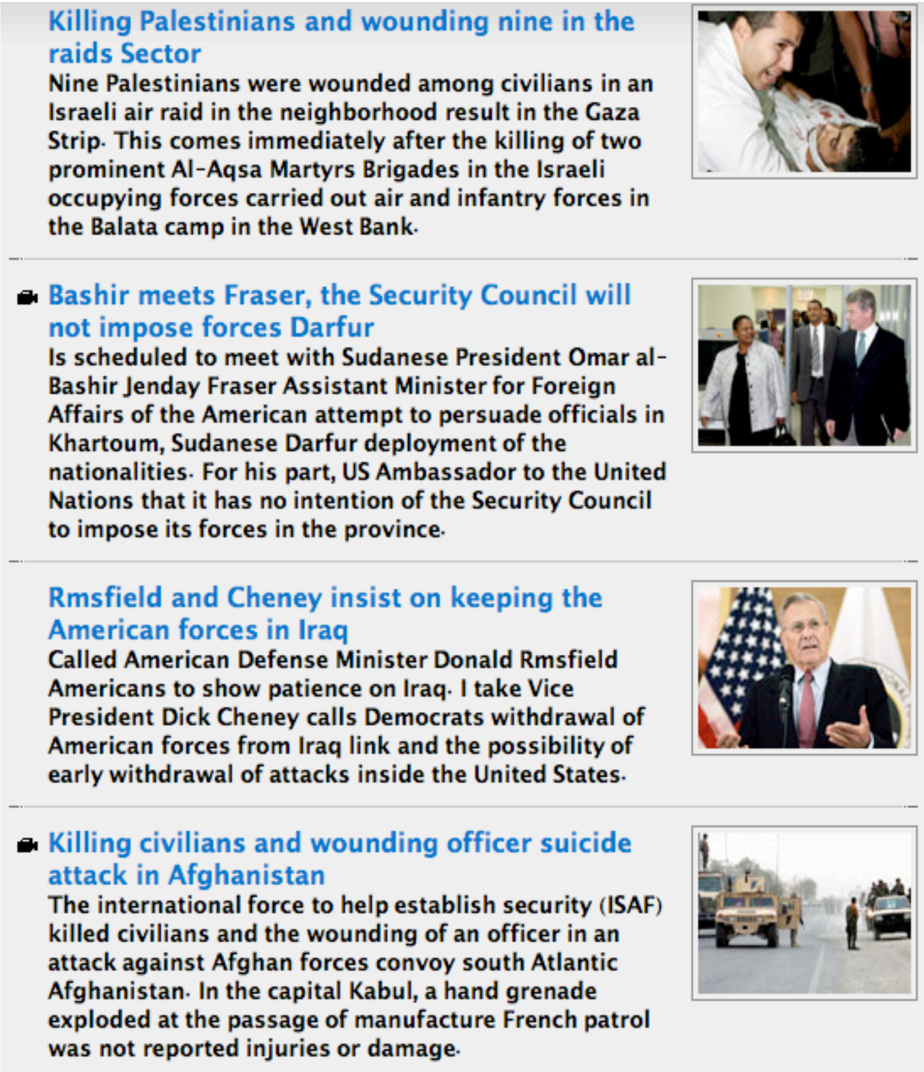
عنوان الحلقة : يوسف ندا.. اتهامات بدعم الإرهاب

الأربعاء/مباشر

19:05 غرينتش 22:05 مكة

frontiers@aljazeera.net

Google Translate



Killing Palestinians and wounding nine in the raids Sector

Nine Palestinians were wounded among civilians in an Israeli air raid in the neighborhood result in the Gaza Strip. This comes immediately after the killing of two prominent Al-Aqsa Martyrs Brigades in the Israeli occupying forces carried out air and infantry forces in the Balata camp in the West Bank.

Bashir meets Fraser, the Security Council will not impose forces Darfur

Is scheduled to meet with Sudanese President Omar al-Bashir Jenday Fraser Assistant Minister for Foreign Affairs of the American attempt to persuade officials in Khartoum, Sudanese Darfur deployment of the nationalities. For his part, US Ambassador to the United Nations that it has no intention of the Security Council to impose its forces in the province.

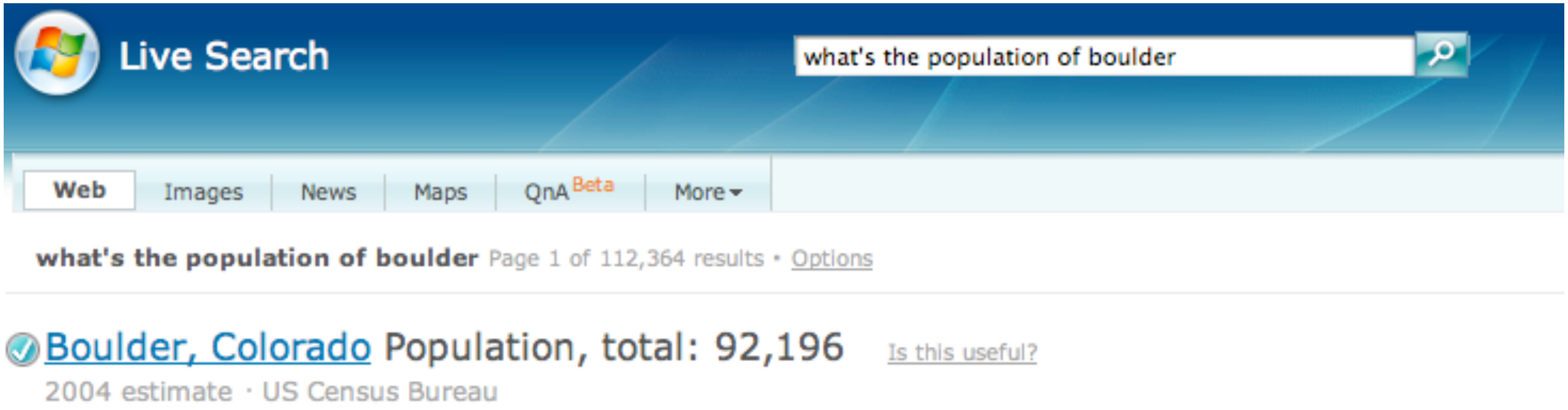
Rmsfield and Cheney insist on keeping the American forces in Iraq

Called American Defense Minister Donald Rmsfield Americans to show patience on Iraq. I take Vice President Dick Cheney calls Democrats withdrawal of American forces from Iraq link and the possibility of early withdrawal of attacks inside the United States.

Killing civilians and wounding officer suicide attack in Afghanistan

The international force to help establish security (ISAF) killed civilians and the wounding of an officer in an attack against Afghan forces convoy south Atlantic Afghanistan. In the capital Kabul, a hand grenade exploded at the passage of manufacture French patrol was not reported injuries or damage.

Web Q/A



what's the population of boulder Page 1 of 112,364 results · Options

✓ Boulder, Colorado Population, total: 92,196 Is this useful?

2004 estimate · US Census Bureau

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).

Web Analytics

Major Topics

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

5. Applications exploiting each

Applications

- First, what makes an application a *language processing application* (as opposed to any other piece of software)?
    - An application that requires the use of knowledge about human languages
        - Example: Is Unix wc (word count) an example of a language processing application?

Applications

- Word count?
  - When it counts words: Yes
    - To count words you need to know what a word is. That's knowledge of language.
  - When it counts lines and bytes: No
    - Lines and bytes are computer artifacts, not linguistic entities

Big Applications

- Question answering
- Conversational agents
- Summarization
- Machine translation

Big Applications

- These kinds of applications require a tremendous amount of knowledge of language.
- Consider the following interaction with HAL the computer from 2001: A Space Odyssey

HAL from 2001

- Dave: *Open the pod bay doors, Hal.*
- HAL: *I'm sorry Dave, I'm afraid I can't do that.*

What’s needed?

- Speech recognition and synthesis
- Knowledge of the English words involved
  - What they mean
- How groups of words clump
  - What the clumps mean

What’s needed?

- Dialog
  - It is polite to respond, even if you’re planning to kill someone.
  - It is polite to pretend to want to be cooperative (I’m afraid, I can’t…)

Caveat

NLP has an AI aspect to it.

- We're often dealing with ill-defined problems
- We don't often come up with exact solutions/algorithms
- We can't let either of those facts get in the way of making progress

Course Material

- We’ll be intermingling discussions of:
  - Linguistic topics
    - E.g. Morphology, syntax, discourse structure
  - Formal systems
    - E.g. Regular languages, context-free grammars
  - Applications
    - E.g. Machine translation, information extraction

Topics: Linguistics

- Word-level processing
- Syntactic processing
- Lexical and compositional semantics
- Discourse processing
- Dialogue structure

Topics: Techniques

- Finite-state methods
- Context-free methods
- Augmented grammars
  - Unification
  - Lambda calculus
- First order logic

- Probability models
- Supervised machine learning methods

Topics: Applications

- Small
  - Spelling correction
  - Hyphenation
- Medium
  - Word-sense disambiguation
  - Named entity recognition
  - Information retrieval
- Large
  - Question answering
  - Conversational agents
  - Machine translation

- Stand-alone
- Enabling applications
- Funding/Business plans

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.

Ambiguity

- Computational linguists are obsessed with ambiguity
- Ambiguity is a fundamental problem of computational linguistics
- Resolving ambiguity is a crucial goal

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.
  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?

Models and Algorithms

- By models we mean the formalisms that are used to capture the various kinds of linguistic knowledge we need.
- Algorithms are then used to manipulate the knowledge representations needed to tackle the task at hand.

Models

- State machines
- Rule-based approaches
- Logical formalisms
- Probabilistic models

Algorithms

- Many of the algorithms that we'll study will turn out to be transducers; algorithms that take one kind of structure as input and output another.
- Unfortunately, ambiguity makes this process difficult. This leads us to employ algorithms that are designed to handle ambiguity of various kinds

Paradigms

- In particular..
  - State-space search
    - To manage the problem of making choices during processing when we lack the information needed to make the right choice
  - Dynamic programming
    - To avoid having to redo work during the course of a state-space search
      - CKY, Earley, Minimum Edit Distance, Viterbi, Baum-Welch
  - Classifiers
    - Machine learning based classifiers that are trained to make decisions based on features extracted from the local context

State Space Search

- States represent pairings of partially processed inputs with partially constructed representations.
- Goals are inputs paired with completed representations that satisfy some criteria.
- As with most interesting problems the spaces are normally too large to exhaustively explore.
  - We need heuristics to guide the search
  - Criteria to trim the space

Dynamic Programming

- Don’t do the same work over and over.
- Avoid this by building and making use of solutions to sub-problems that must be invariant across all parts of the space.