Lecture 38 of 42

Natural Language Processing, Part 1: Machine Translation

William H. Hsu

Department of Computing and Information Sciences, KSU

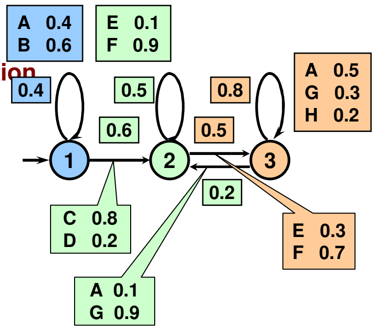
KSOL course page: http://snipurl.com/v9v3
Course web site: http://www.kddresearch.org/Courses/CIS730
Instructor home page: http://www.cis.ksu.edu/~bhsu

Reading for Next Class:

Chapter 22.4 – 22.9, p. 806 – 826, Russell and Norvig

Learning Framework for Natural Language: (Hidden) Markov Models

- **Definition of Hidden Markov Models (HMMs)**
  - **Stochastic state transition diagram (HMMs: states, *aka* nodes, are hidden)**
  - **Compare: probabilistic finite state automaton (Mealy/Moore model)**
  - **Annotated transitions (*aka* arcs, edges, links)**
    - ***Output alphabet* (the observable part)**
    - **Probability distribution over *outputs***
- **Forward Problem: One Step in ML Estimation**
  - **Given: model *h*, observations (data) *D***
  - **Estimate: $P(D \mid h)$**
- **Backward Problem: Prediction Step**
  - **Given: model *h*, observations *D***
  - **Maximize: $P(h(X) = x \mid h, D)$ for a new *X***
- **Forward-Backward (Learning) Problem**
  - **Given: model space *H*, data *D***
  - **Find: $h \in H$ such that $P(h \mid D)$ is maximized (i.e., MAP hypothesis)**
- **HMMs Also A Case of LSQ (*f* Values in [Roth, 1999])**



NLP Issues:
Word Sense Disambiguation (WSD)

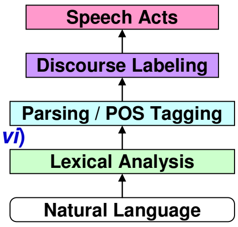
- **Problem Definition**
  - Given: *m* sentences, each containing a usage of a particular ambiguous word
  - Example: "The can will rust." (auxiliary verb versus noun)
  - Label: $v_j \equiv s \equiv$ correct word sense (e.g., $s \in$ {auxiliary verb, noun})
  - Representation: *m* examples (labeled attribute vectors $<(w_1, w_2, \ldots, w_n), s>$)
  - Return: classifier $f: X \rightarrow V$ that disambiguates new $x \equiv (w_1, w_2, \ldots, w_n)$
- **Solution Approach: Use Bayesian Learning (e.g., Naïve Bayes)**
  - Caveat: can't observe *s* in the text!
  - A solution: treat *s* in $P(w_i \mid s)$ as *missing value*, impute *s* (assign by inference)
  - [Pedersen and Bruce, 1998]: fill in using Gibbs sampling, EM algorithm (later)
  - [Roth, 1998]: Naïve Bayes, sparse networks of Winnows (SNOW), TBL

$$P(w_1, w_2, \ldots, w_n / s) = \prod_{i=1}^{n} P(w_i / s)$$

- **Recent Research**
  - T. Pedersen's research home page: http://www.d.umn.edu/~tpederse/
  - D. Roth's Cognitive Computation Group: http://l2r.cs.uiuc.edu/~cogcomp/

NLP Issues:
Part-of-Speech (POS) Tagging

- **Problem Definition**
  - Given: *m* sentences containing untagged words
  - Example: "The can will rust."
  - Label (one per word, out of ~30-150): $v_j \equiv s \equiv$ (*art*, *n*, *aux*, *vi*)
  - Representation: labeled examples $<(w_1, w_2, \ldots, w_n), s>$
  - Return: classifier $f: X \rightarrow V$ that tags $x \equiv (w_1, w_2, \ldots, w_n)$
  - Applications: WSD, dialogue acts (e.g., "That sounds OK to me." → *ACCEPT*)



- **Solution Approaches: Use Transformation-Based Learning (TBL)**
  - [Brill, 1995]: TBL - mistake-driven algorithm that produces sequences of rules
    - Each rule of form (t_i, v): a test condition (constructed attribute) and tag
    - t_i: "*w* within ±*k* words of w_i" (context words); collocations (windows)
  - For more info: see [Roth, 1998], [Samuel, Carberry, Vijay-Shankar, 1998]
- **Recent Research**
  - E. Brill's page: http://www.cs.jhu.edu/~brill/
  - K. Samuel's page: http://www.eecis.udel.edu/~samuel/work/research.html

NLP Applications: Info Retrieval (IR) and Digital Libraries

- Information Retrieval (IR)
  - One role of learning: produce classifiers for documents (see [Sahami, 1999])
  - Query-based search engines (e.g., for WWW: *AltaVista*, *Lycos*, *Yahoo*)
  - Applications: bibliographic searches (citations, patent intelligence, etc.)
- Bayesian Classification: Integrating Supervised and Unsupervised Learning
  - Unsupervised learning: organize collections of documents at a “topical” level
  - e.g., *AutoClass* [Cheeseman *et al*, 1988]; self-organizing maps [Kohonen, 1995]
  - More on this topic (document clustering) soon
- Framework Extends Beyond Natural Language
  - Collections of images, audio, video, other media
  - Five *Ss* : Source, Stream, Structure, Scenario, Society
  - Book on IR [vanRijsbergen, 1979]: http://www.dcs.gla.ac.uk/Keith/Preface.html
- Recent Research
  - M. Sahami’s page (Bayesian IR): http://robotics.stanford.edu/users/sahami
  - Digital libraries (DL) resources: http://fox.cs.vt.edu

Statistical Machine Translation

Kevin Knight

USC/Information Sciences Institute
USC/Computer Science Department

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

➡

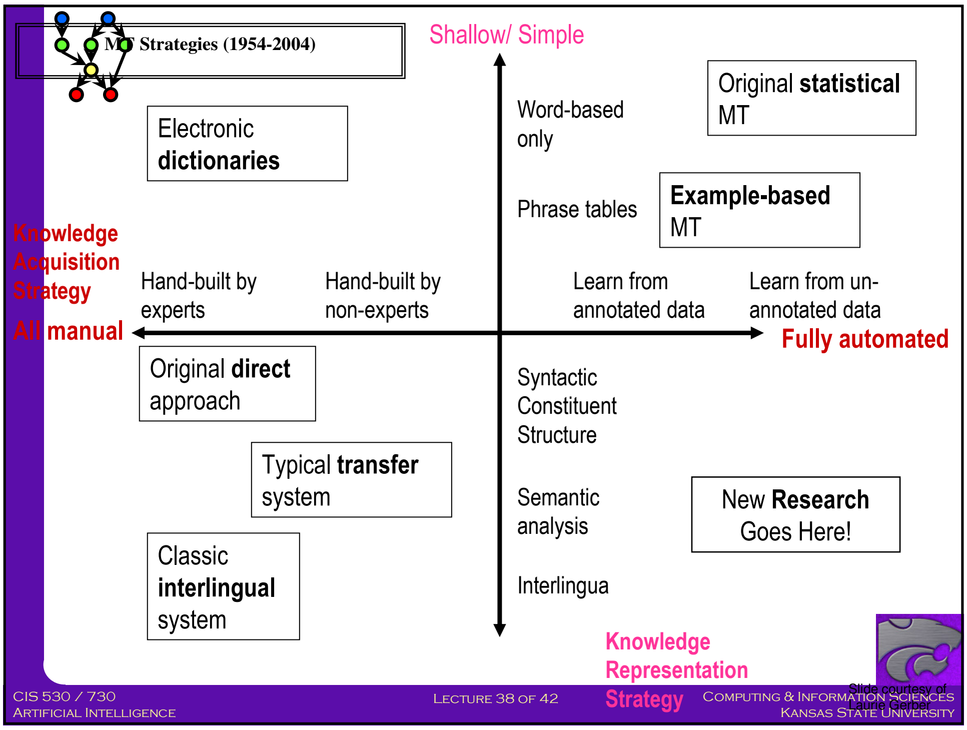
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

The classic acid test for natural language processing.

Requires capabilities in both interpretation and generation.

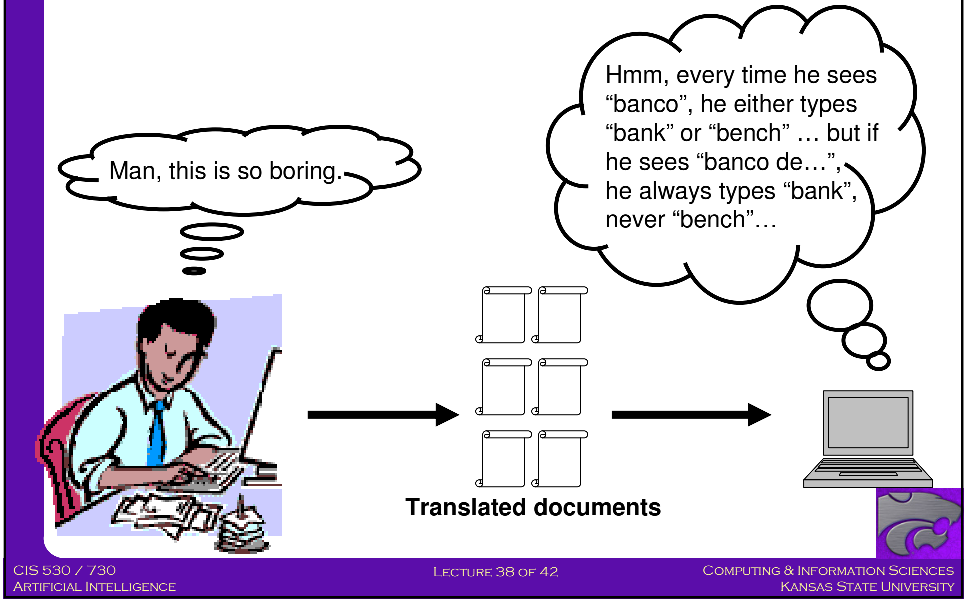
About $10 billion spent annually on human translation.

MT Strategies (1954-2004)


Shallow/ Simple
Word-based only
Original **statistical** MT
Electronic **dictionaries**
Phrase tables
Example-based MT
Knowledge Acquisition Strategy
Hand-built by experts
Hand-built by non-experts
Learn from annotated data
Learn from un-annotated data
All manual
Fully automated
Original **direct** approach
Syntactic Constituent Structure
Typical **transfer** system
Semantic analysis
New **Research** Goes Here!
Classic **interlingual** system
Interlingua
Knowledge Representation Strategy

Slide courtesy of Laurie Gerber

Data-Driven Machine Translation


Man, this is so boring.
Hmm, every time he sees "banco", he either types "bank" or "bench" … but if he sees "banco de…", he always types "bank", never "bench"…
Translated documents

Recent Progress in Statistical MT

slide from C. Wayne, DARPA

2002

insistent Wednesday may recurred her trips to Libya tomorrow for flying

Cairo 6-4 (AFP) - an official announced today in the Egyptian lines company for flying Tuesday is a company " insistent for flying " may resumed a consideration of a day Wednesday tomorrow her trips to Libya of Security Council decision trace international the imposed ban comment .

And said the official " the institution sent a speech to Ministry of Foreign Affairs of lifting on Libya air , a situation her receiving replying are so a trip will pull to Libya a morning Wednesday " .

2003

Egyptair Has Tomorrow to Resume Its Flights to Libya

Cairo 4-6 (AFP) - said an official at the Egyptian Aviation Company today that the company egyptair may resume as of tomorrow, Wednesday its flights to Libya after the International Security Council resolution to the suspension of the embargo imposed on Libya.

" The official said that the company had sent a letter to the Ministry of Foreign Affairs, information on the lifting of the air embargo on Libya, where it had received a response, the first take off a trip to Libya on Wednesday morning ".

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .

3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok **farok** izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok **farok** ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok clok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .

3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok **farok** izok stok .

5b. totat **jjat** quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok **farok** ororok lalok sprok izok enemok .

7b. wat **jjat** bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok clok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .

3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok **farok** izok stok .

/

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok **farok** ororok lalok sprok izok enemok .

/

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok clok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok **crrrok** hihok yorok zanzanok .

???

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok **hihok** ghirok .

3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

/

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

/

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok clok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok **hihok** yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok **hihok** mok .

12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok** yorok clok kantok ok-yurp

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok **hihok** ghirok .

3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok **yorok** ghirok clok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok **hihok** **yorok** zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok **hihok** mok .

12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .

3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anok plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok **clok** .

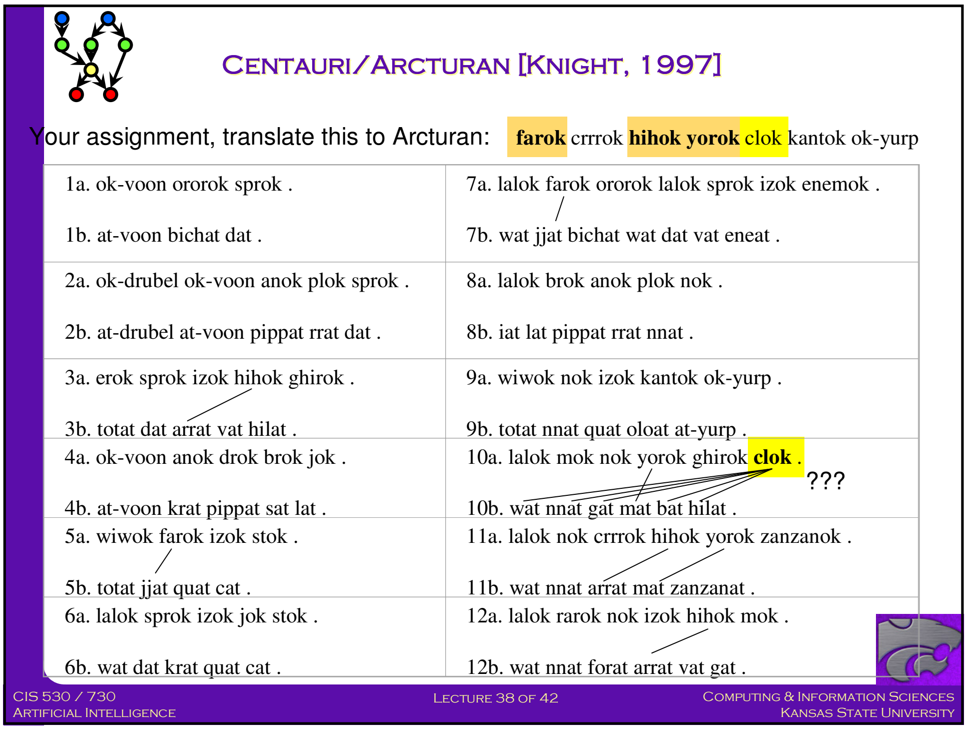
10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .



Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok .
1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .
2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .
3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .
4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .
5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .
6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .
7b. wat jjat bichat wat dat vat eneat .

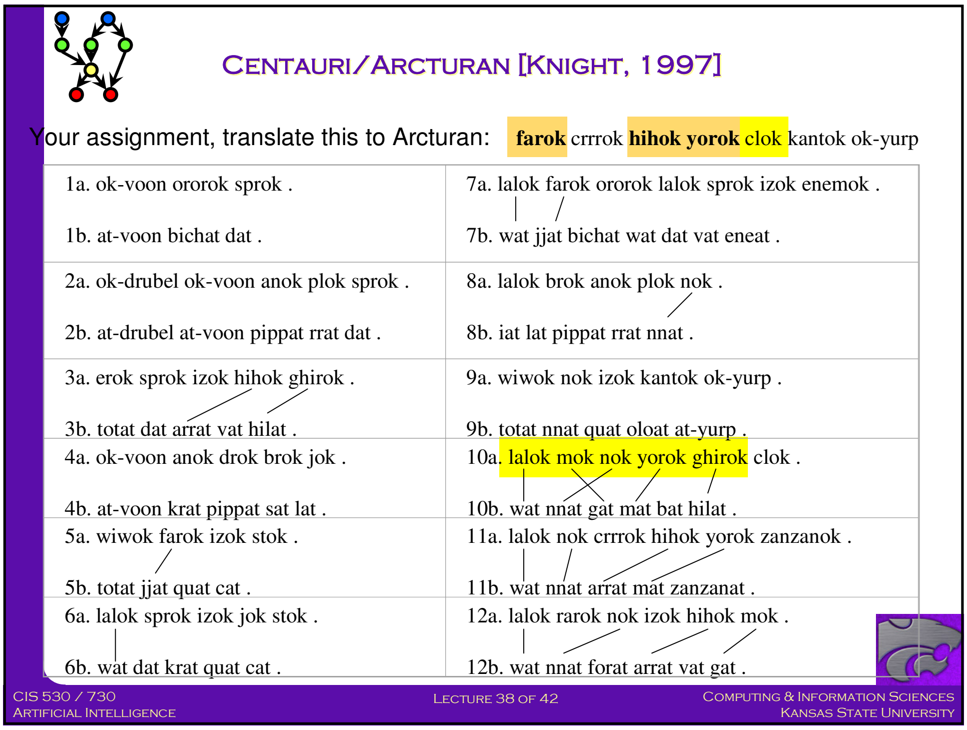
8a. lalok brok anok plok nok .
8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .
9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok **clok** .
???
10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .
11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .
12b. wat nnat forat arrat vat gat .



Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok .
1b. at-voon bichat dat .

2a. ok-drubel ok-voon anok plok sprok .
2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghirok .
3b. totat dat arrat vat hilat .

4a. ok-voon anok drok brok jok .
4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .
5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .
6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .
7b. wat jjat bichat wat dat vat eneat .

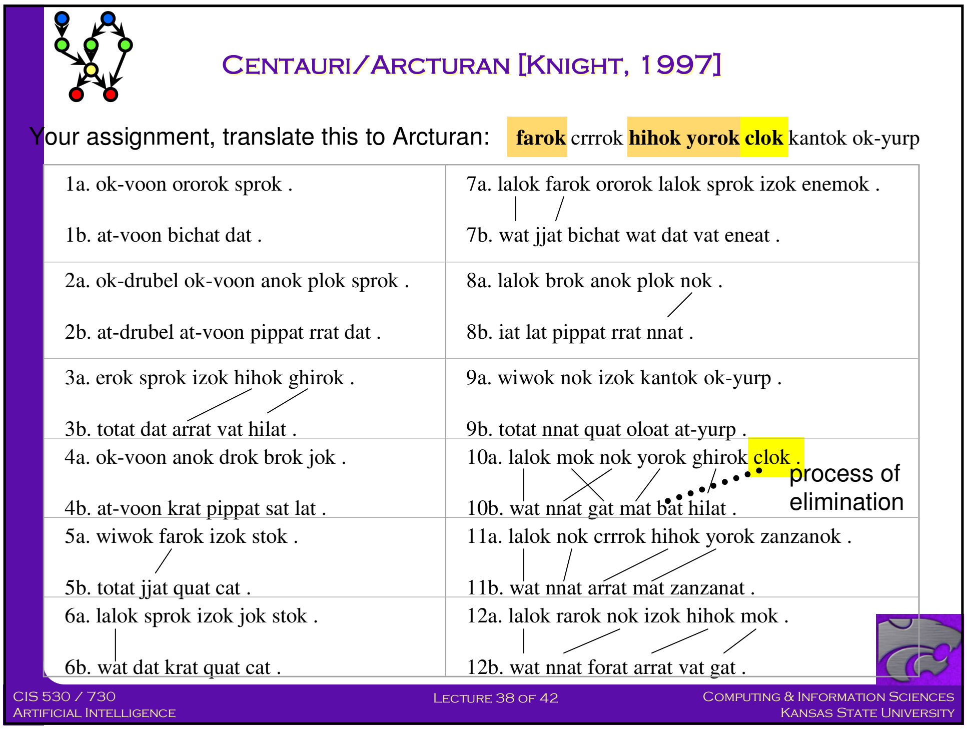
8a. lalok brok anok plok nok .
8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .
9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghirok clok .
10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .
11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .
12b. wat nnat forat arrat vat gat .

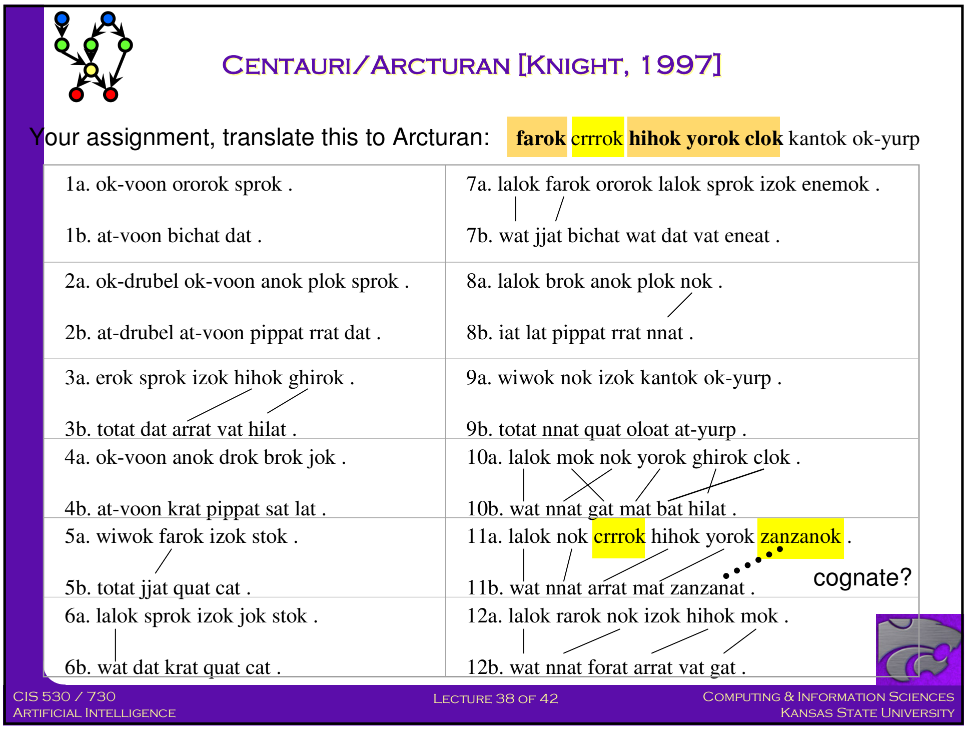


Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** **clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

process of elimination

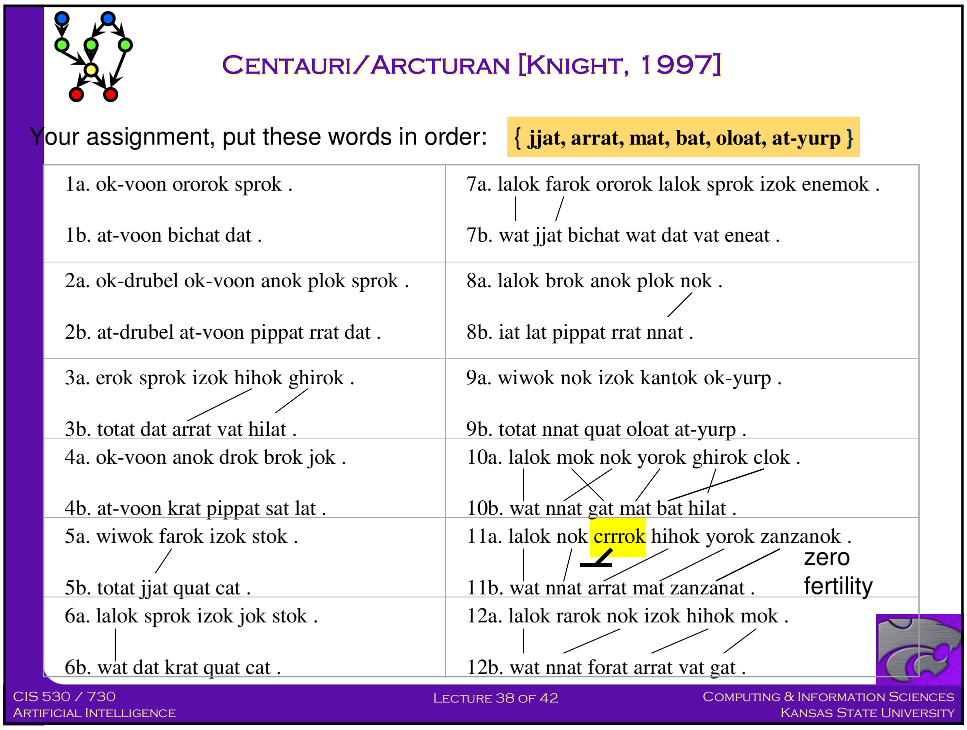


Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

cognate?

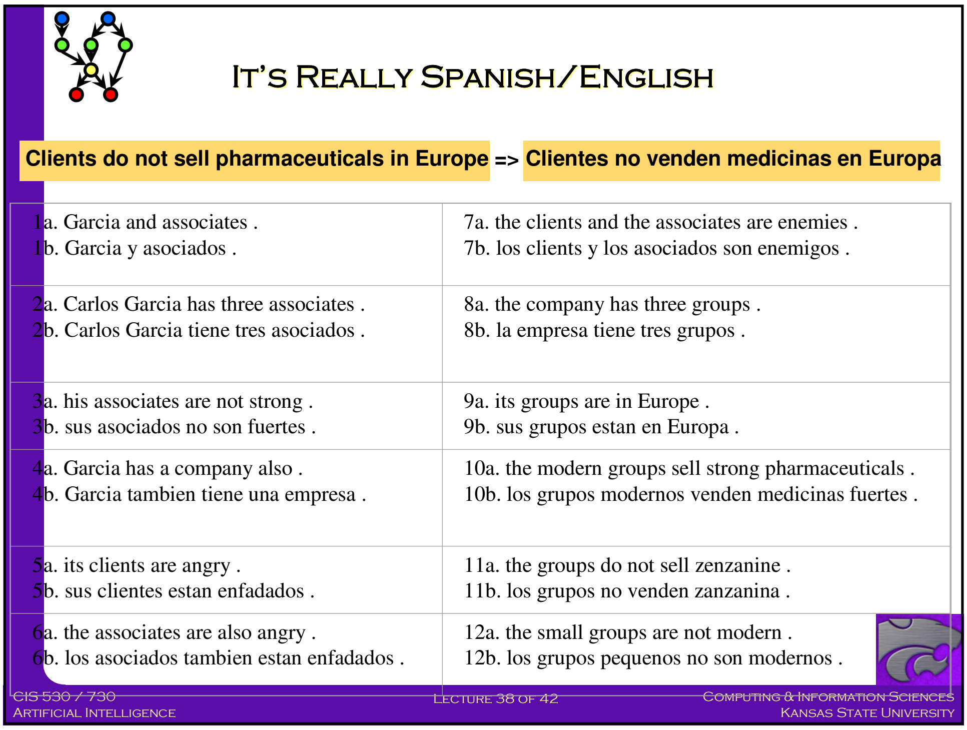


Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order: **{ jjat, arrat, mat, bat, oloat, at-yurp }**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

zero fertility



It's Really Spanish/English

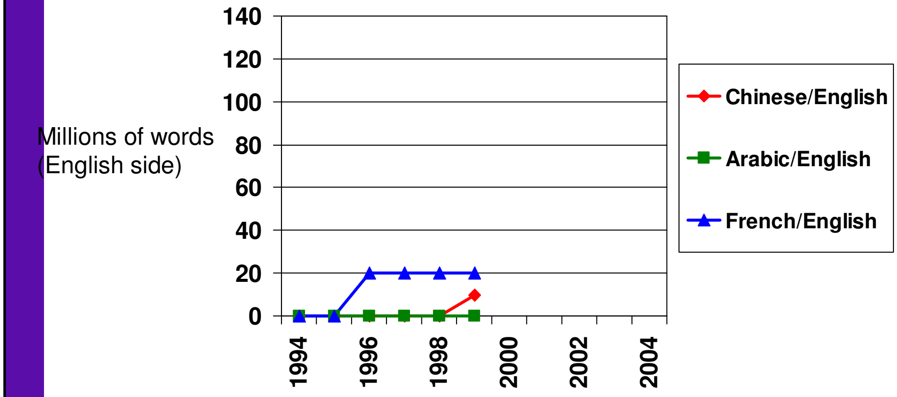
Clients do not sell pharmaceuticals in Europe => **Clientes no venden medicinas en Europa**

1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

Data for Statistical MT

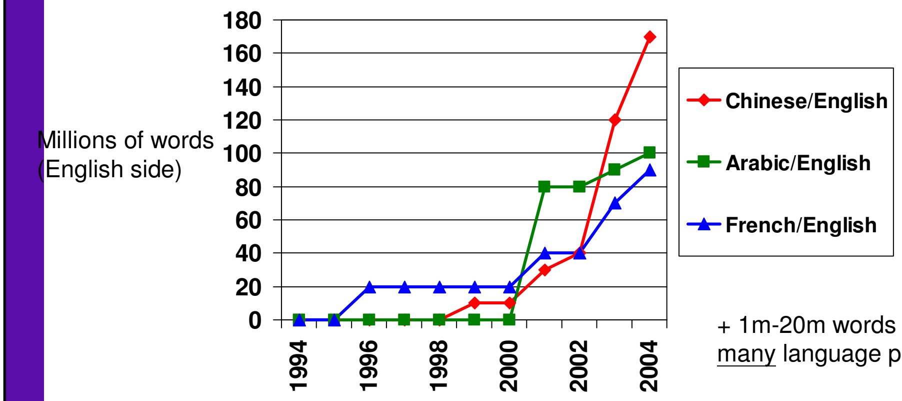
and data preparation

Ready-to-Use Online Bilingual Data



(Data stripped of formatting, in sentence-pair format, available from the Linguistic Data Consortium at UPenn).

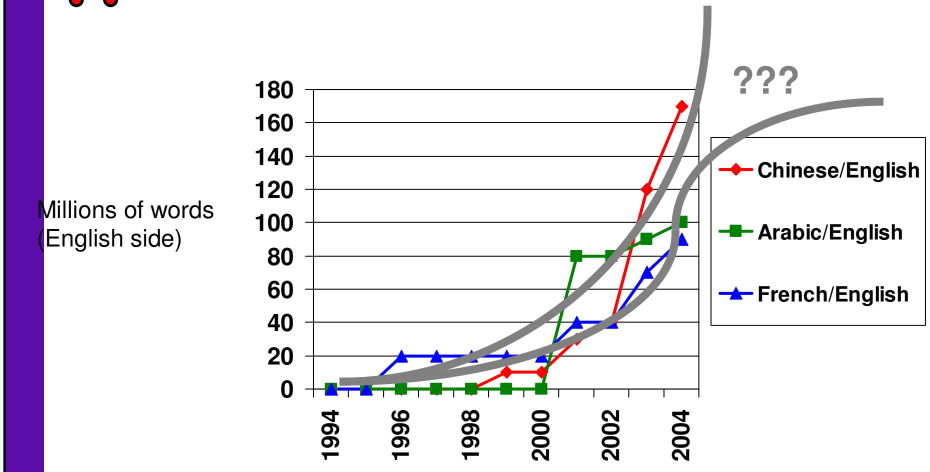
Ready-to-Use Online Bilingual Data



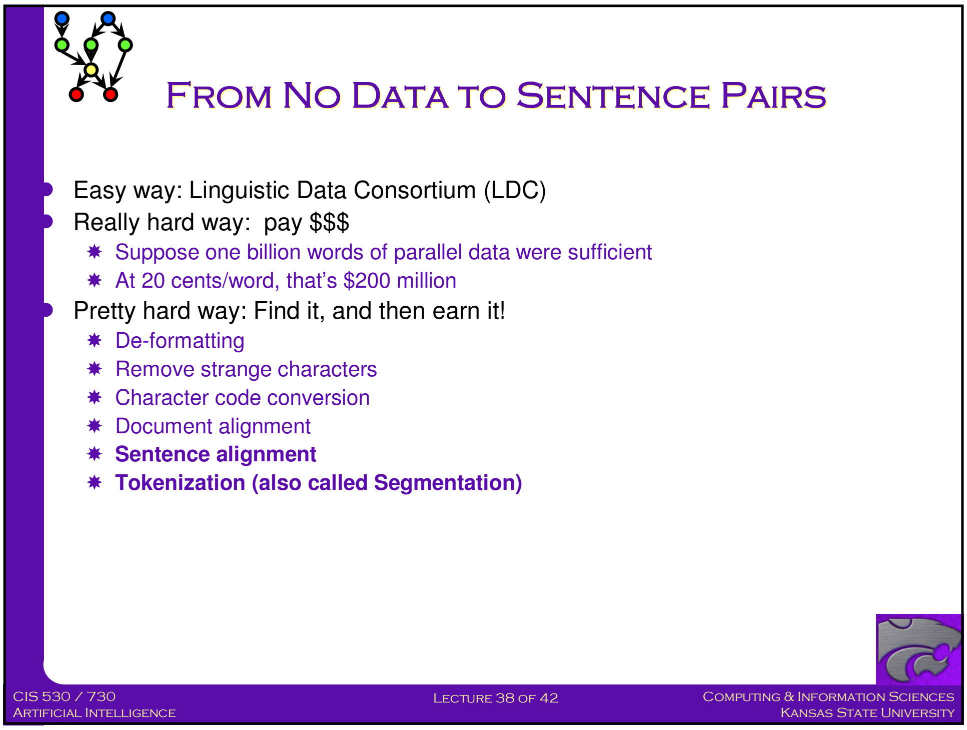
+ 1m-20m words for many language pairs

(Data stripped of formatting, in sentence-pair format, available from the Linguistic Data Consortium at UPenn).

Ready-to-Use Online Bilingual Data

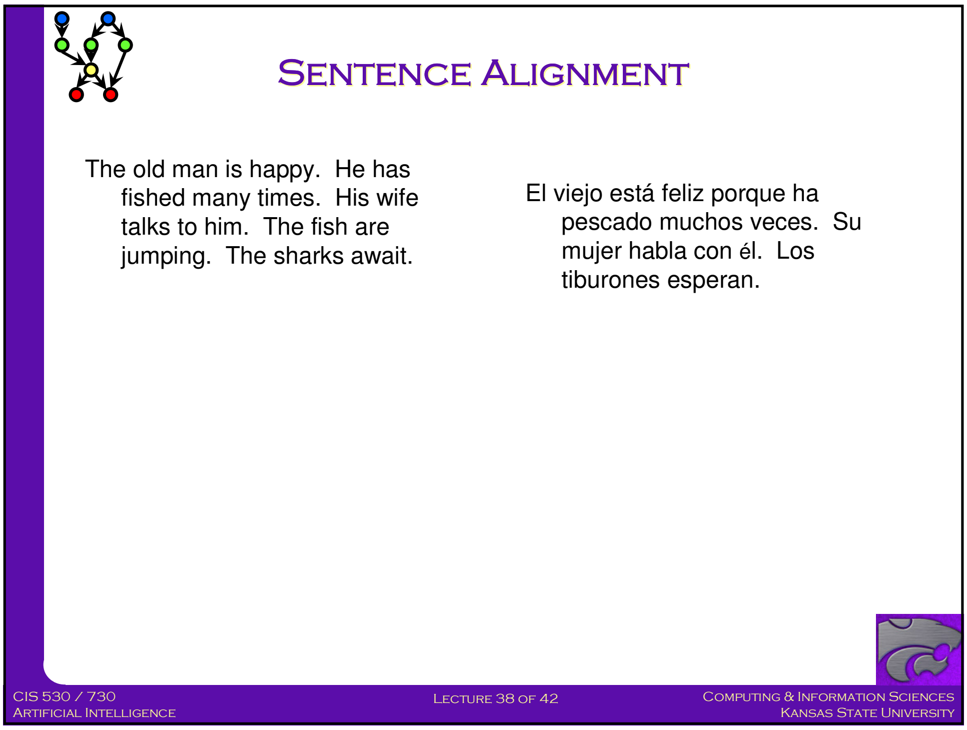


→ One Billion?



From No Data to Sentence Pairs

- Easy way: Linguistic Data Consortium (LDC)
- Really hard way: pay $$$
  - Suppose one billion words of parallel data were sufficient
  - At 20 cents/word, that's $200 million
- Pretty hard way: Find it, and then earn it!
  - De-formatting
  - Remove strange characters
  - Character code conversion
  - Document alignment
  - **Sentence alignment**
  - **Tokenization (also called Segmentation)**



Sentence Alignment

The old man is happy. He has fished many times. His wife talks to him. The fish are jumping. The sharks await.

El viejo está feliz porque ha pescado muchos veces. Su mujer habla con él. Los tiburones esperan.

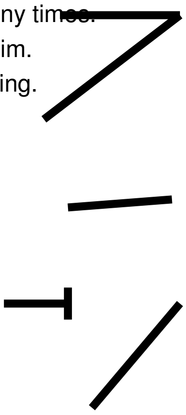
Sentence Alignment

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Sentence Alignment

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.



1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Sentence Alignment

1. The old man is happy. He has fished many times.
2. His wife talks to him.
3. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Note that unaligned sentences are thrown out, and sentences are merged in n-to-m alignments (n, m > 0).

Tokenization (or Segmentation)

- English
  - Input (some byte stream):

```
"There," said Bob.
```

  - Output (7 "tokens" or "words"):

```
" There , " said Bob .
```

- Chinese
  - Input (byte stream):
  - Output:

美国关岛国际机场及其办公室均接获
一名自称沙地阿拉伯富商拉登等发出
的电子邮件。

美国 关岛国 际机 场 及其 办公
室均接获 一名 自称 沙地 阿拉 伯
富 商拉登 等发 出 的 电子邮件。

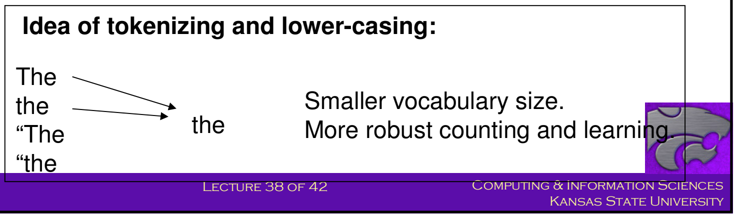
LOWER-CASING

- English
  - Input (7 words):

    " There , " said Bob .

  - Output (7 words):

    " there , " said bob .



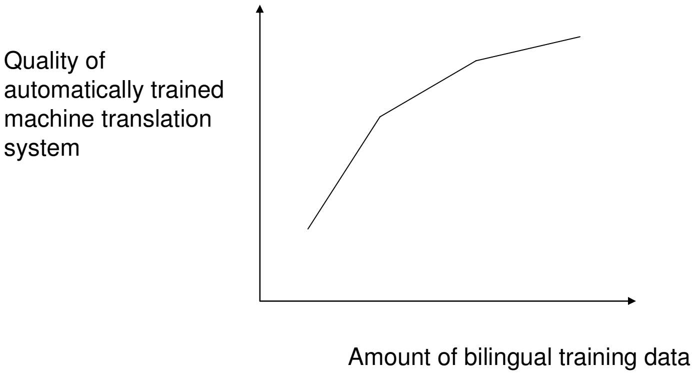
Idea of tokenizing and lower-casing:

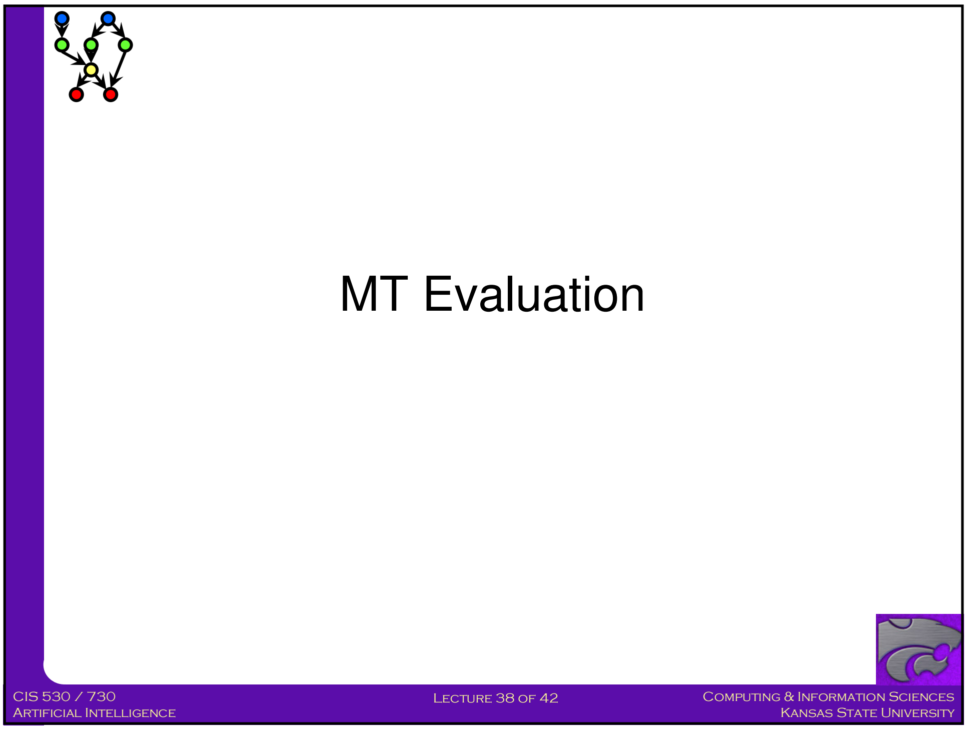
The
the
“The
“the

the

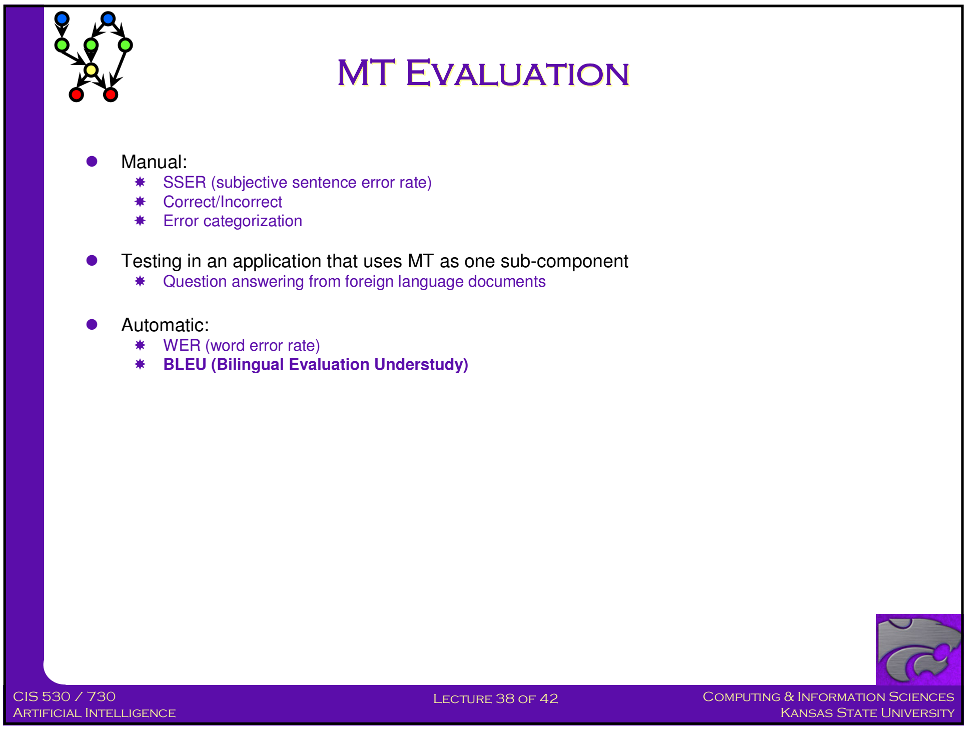
Smaller vocabulary size.
More robust counting and learning

IT IS POSSIBLE TO DRAW LEARNING CURVES: HOW MUCH DATA DO WE NEED?





MT Evaluation

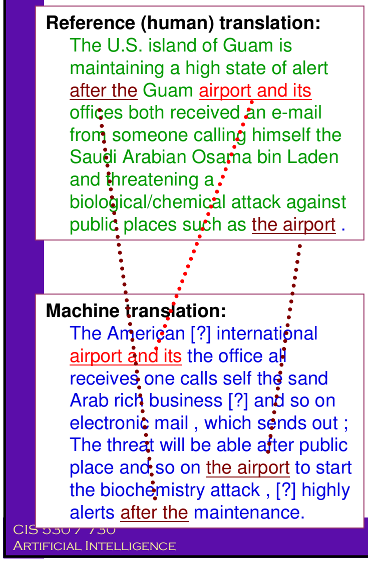


MT Evaluation

- Manual:
  - SSER (subjective sentence error rate)
  - Correct/Incorrect
  - Error categorization
- Testing in an application that uses MT as one sub-component
  - Question answering from foreign language documents
- Automatic:
  - WER (word error rate)
  - **BLEU (Bilingual Evaluation Understudy)**

BLEU Evaluation Metric

(Papineni et al, ACL-2002)



Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

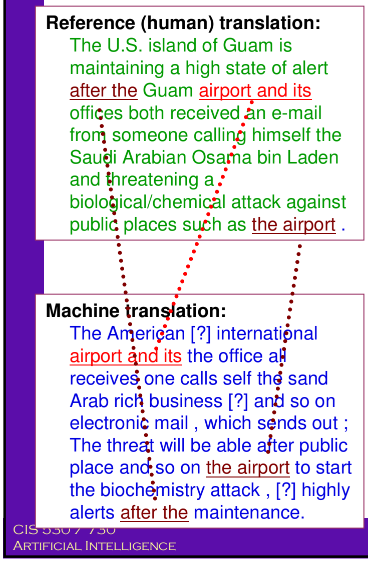
Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
    - An n-gram is an sequence of n words
  - Not allowed to use same portion of reference translation twice (can't cheat by typing out "the the the the the")
- Brevity penalty
  - Can't just type out single word "the" (precision 1.0!)

*** Amazingly hard to "game" the system (i.e., find a way to change machine output so that BLEU goes up, but quality doesn't)

BLEU Evaluation Metric

(Papineni et al, ACL-2002)



Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

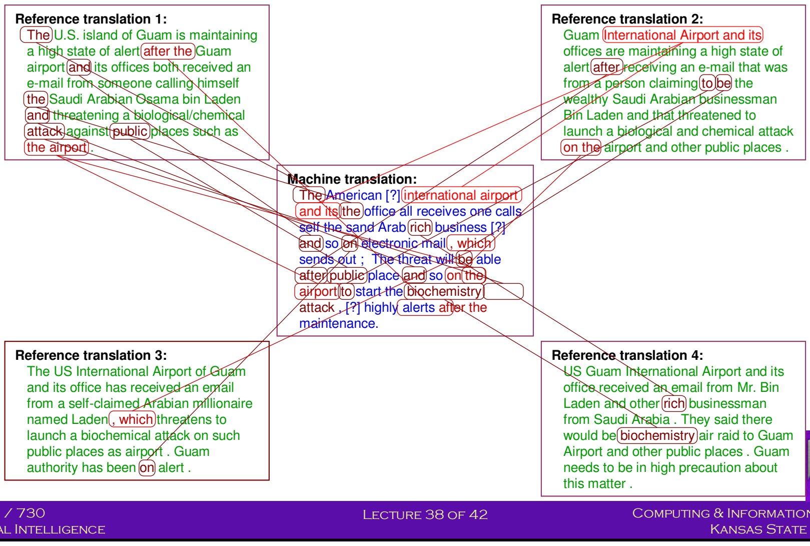
Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- BLEU4 formula
  (counts n-grams up to length 4)

$$\begin{aligned}&\exp(1.0 * \log p1 + \\ &\quad 0.5 * \log p2 + \\ &\quad 0.25 * \log p3 + \\ &\quad 0.125 * \log p4 - \\ &\quad \max(\text{words-in-reference} / \text{words-in-machine} - 1, \\ &\qquad 0)\end{aligned}$$

p1 = 1-gram precision
P2 = 2-gram precision
P3 = 3-gram precision
P4 = 4-gram precision

Multiple Reference Translations



Reference translation 1:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

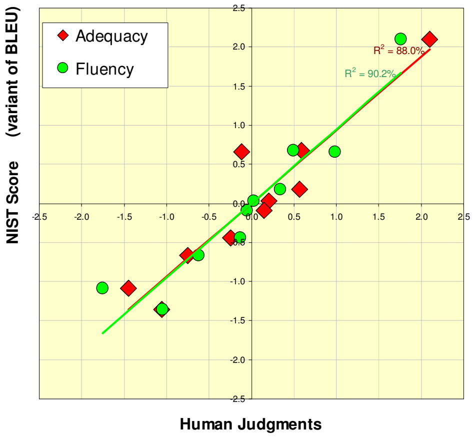
Reference translation 2:
Guam International Airport and its offices are maintaining a high state of alert after receiving an e-mail that was from a person claiming to be the wealthy Saudi Arabian businessman Bin Laden and that threatened to launch a biological and chemical attack on the airport and other public places .

Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

Reference translation 3:
The US International Airport of Guam and its office has received an email from a self-claimed Arabian millionaire named Laden , which threatens to launch a biochemical attack on such public places as airport . Guam authority has been on alert .

Reference translation 4:
US Guam International Airport and its office received an email from Mr. Bin Laden and other rich businessman from Saudi Arabia . They said there would be biochemistry air raid to Guam Airport and other public places . Guam needs to be in high precaution about this matter .

BLEU Tends to Predict Human Judgments



slide from G. Doddington (NIST)

BLEU in Action

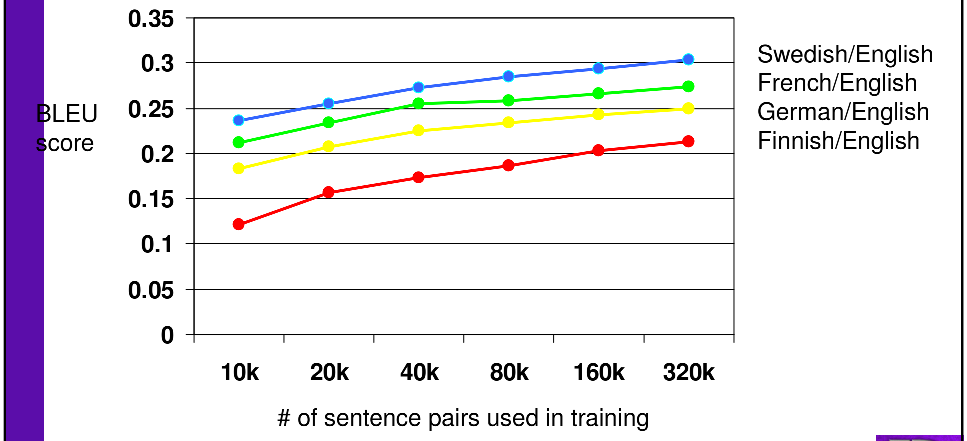
枪手被警方击毙。 **(Foreign Original)**

the gunman was shot to death by the police . **(Reference Translation)**

the gunman was police kill . **#1**
wounded police jaya of **#2**
the gunman was shot dead by the police . **#3**
the gunman arrested by police kill . **#4**
the gunmen were killed . **#5**
the gunman was shot to death by the police . **#6**
gunmen were killed by police ?SUB>0 ?SUB>0 **#7**
al by the police . **#8**
the ringer is killed by the police . **#9**
police killed the gunman . **#10**

green = 4-gram match (good!)
red = word not matched (bad!)

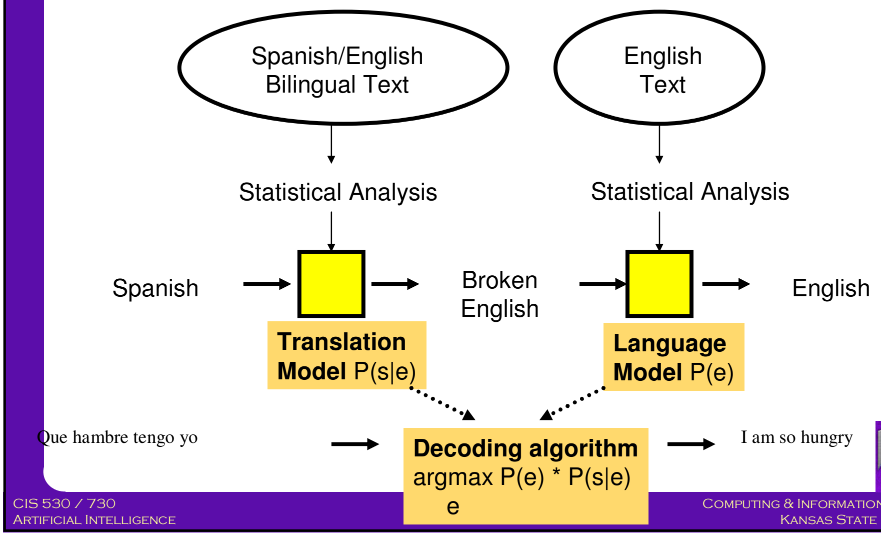
Sample Learning Curves

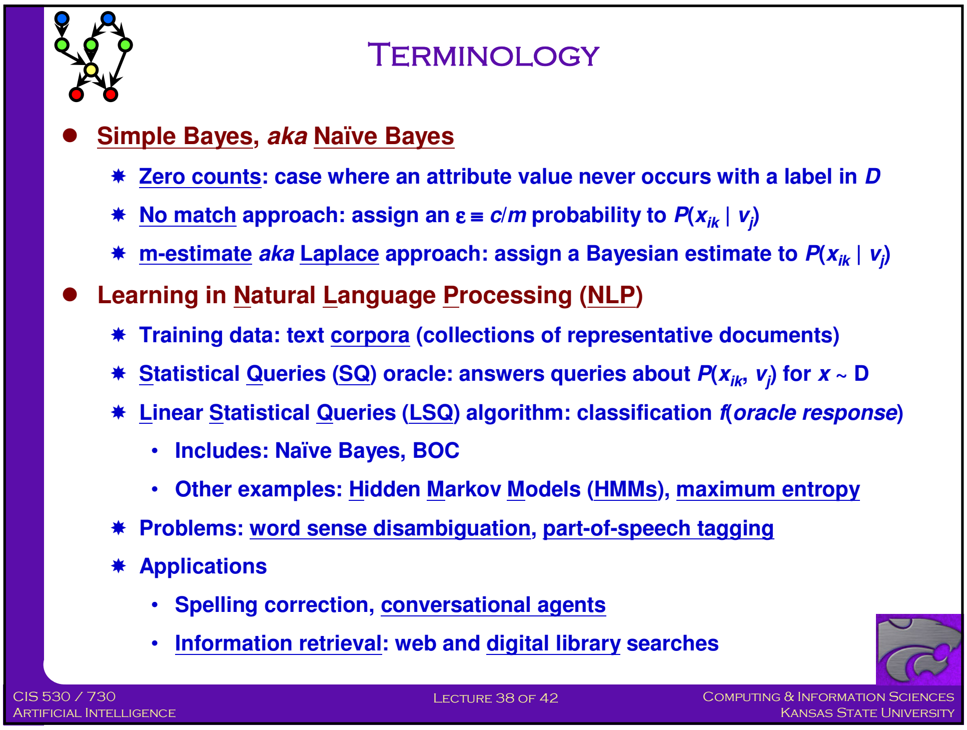


Experiments by
Philipp Koehn

Word-Based Statistical MT

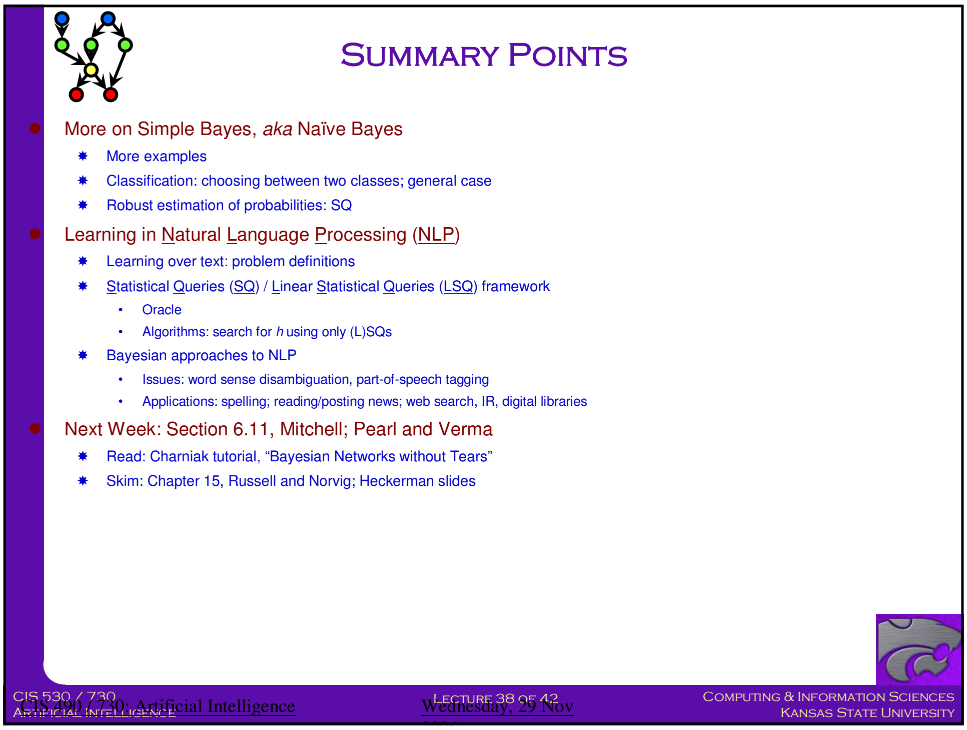
Statistical MT Systems





Terminology

- **Simple Bayes, *aka* Naïve Bayes**
  - Zero counts: case where an attribute value never occurs with a label in *D*
  - No match approach: assign an ε ≡ c/m probability to $P(x_{ik} \mid v_j)$
  - m-estimate *aka* Laplace approach: assign a Bayesian estimate to $P(x_{ik} \mid v_j)$
- **Learning in Natural Language Processing (NLP)**
  - Training data: text corpora (collections of representative documents)
  - Statistical Queries (SQ) oracle: answers queries about $P(x_{ik}, v_j)$ for $x \sim$ D
  - Linear Statistical Queries (LSQ) algorithm: classification *f*(*oracle response*)
    - Includes: Naïve Bayes, BOC
    - Other examples: Hidden Markov Models (HMMs), maximum entropy
  - Problems: word sense disambiguation, part-of-speech tagging
  - Applications
    - Spelling correction, conversational agents
    - Information retrieval: web and digital library searches



Summary Points

- More on Simple Bayes, *aka* Naïve Bayes
  - More examples
  - Classification: choosing between two classes; general case
  - Robust estimation of probabilities: SQ
- Learning in Natural Language Processing (NLP)
  - Learning over text: problem definitions
  - Statistical Queries (SQ) / Linear Statistical Queries (LSQ) framework
    - Oracle
    - Algorithms: search for *h* using only (L)SQs
  - Bayesian approaches to NLP
    - Issues: word sense disambiguation, part-of-speech tagging
    - Applications: spelling; reading/posting news; web search, IR, digital libraries
- Next Week: Section 6.11, Mitchell; Pearl and Verma
  - Read: Charniak tutorial, “Bayesian Networks without Tears”
  - Skim: Chapter 15, Russell and Norvig; Heckerman slides