

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Machine Translation

Many slides courtesy of Dan Klein, Stuart Russell, or Andrew Moore

CS 5300 / CS 6300
Artificial Intelligence
Spring 2009

Hal Daumé III
hal@cs.utah.edu

www.cs.utah.edu/~hal/courses/2009S_AI

Slide 1 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Why Study Language?

It is the fundamental method of human communication.

\$ms are spent paying people to process language each year.

It constantly surprises us.

There's a lot of it!

Slide 2 CS 5300: MT

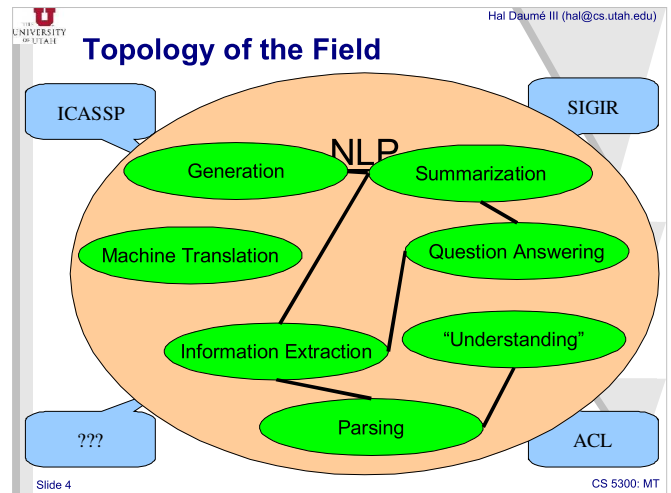
UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Translate Centauri -> Arcturan

Your assignment, translate this Centauri sentence to Arcturan:
farok crrok hihok yorok klok 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 anak plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anak plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghrok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anak drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghrok klok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrok hihok yorok zanzanak . 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 .

Slide 3 CS 5300: MT

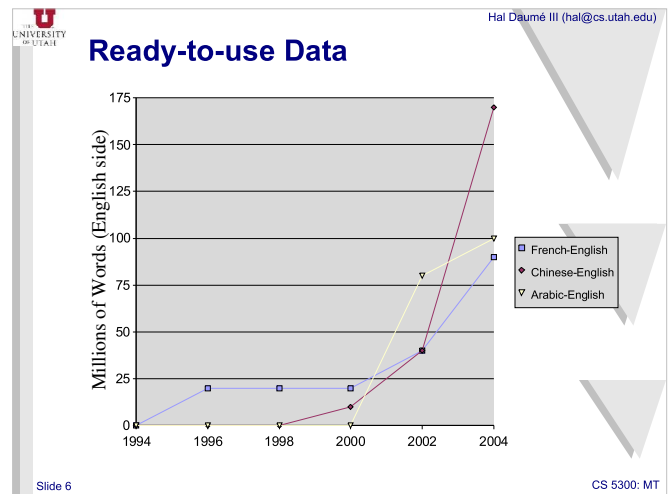


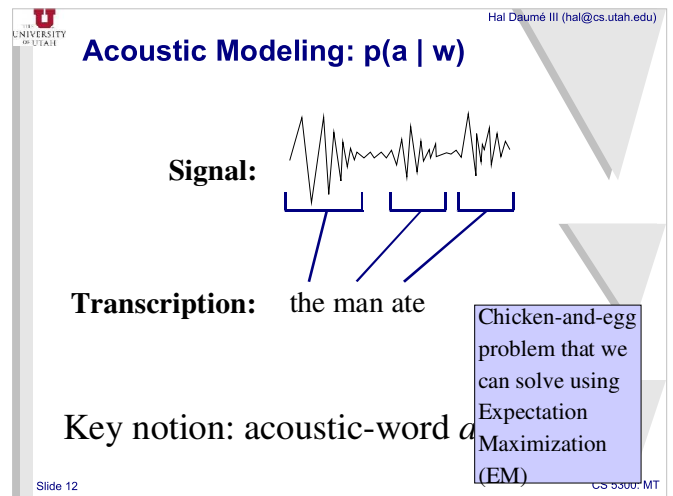
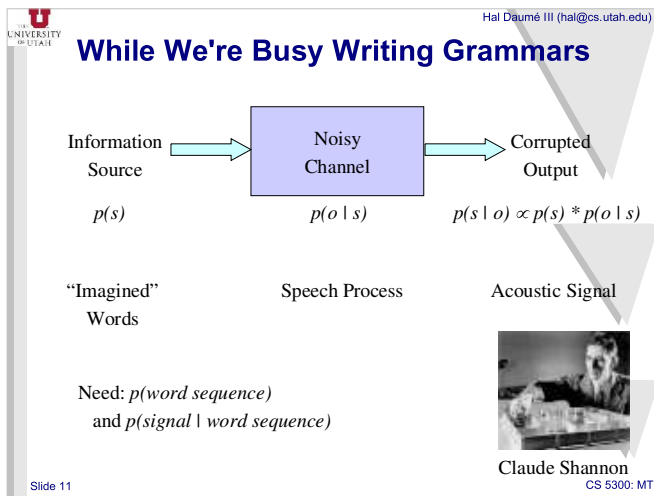
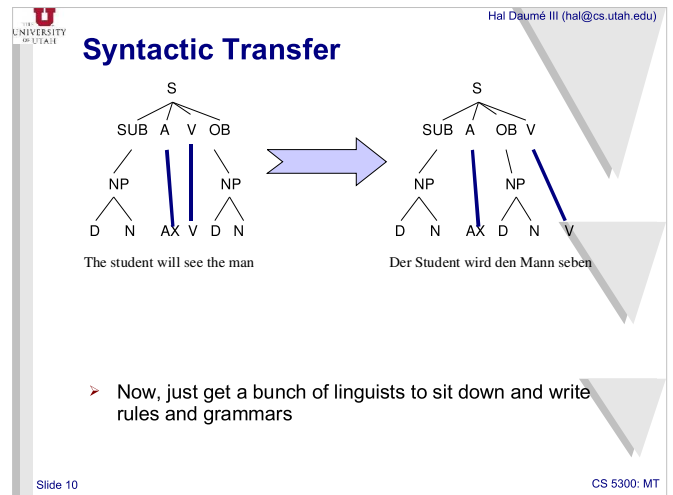
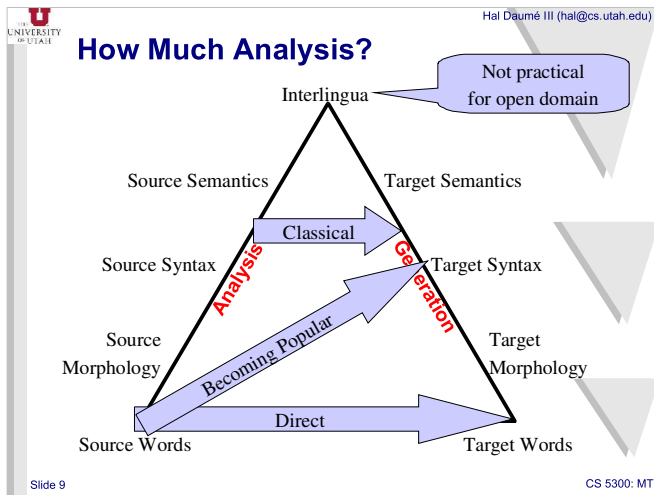
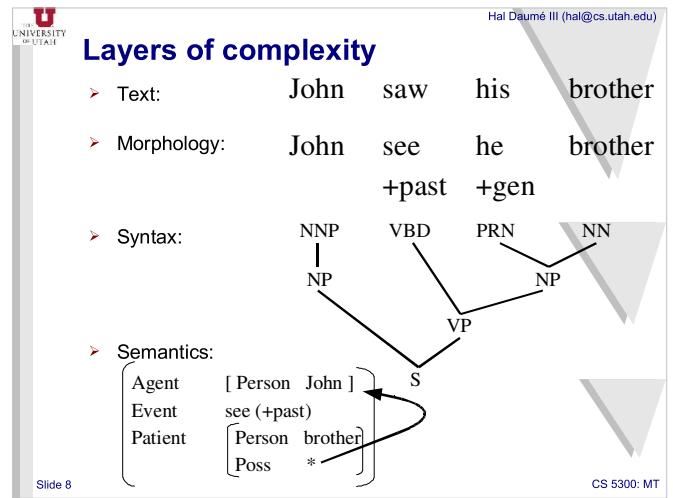
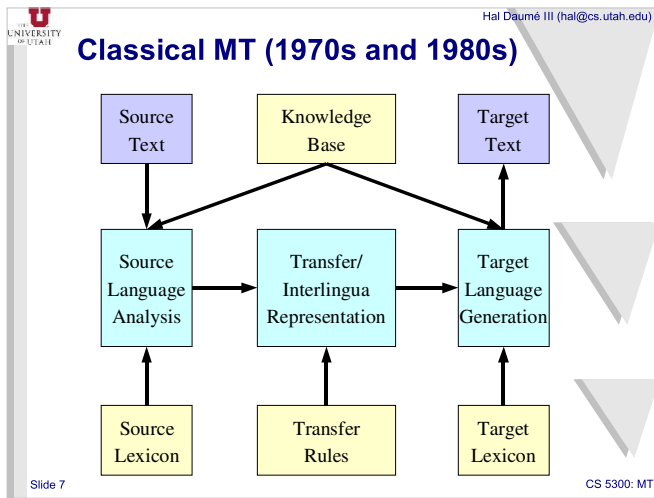
UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

A Bit of History

- 1940s** Computations begins, AI hot, Turing test
Machine translation = Code-breaking?
- 1950s** Cold war continues
- 1960s** Chomsky and statistics, ALPAC report
- 1970s** Dry spell
- 1980s** Statistics makes significant advances in speech
- 1990s** Web arrives
Statistical revolution in machine translation, parsing, IE, etc
Serious "corpus" work, increasing focus on evaluation
- 2000s** Focus on optimizing loss functions, reranking
How much can we automate?
Huge process in machine translation
Gigantic corpora become available, scaling
New challenges

Slide 5 CS 5300: MT





UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Language Modeling: $p(w)$

- Sentence = sequence of symbols from an alphabet

$$p(w_1, w_2, \dots, w_I) = \prod_{i=1}^I p(w_i | w_1, \dots, w_{i-1})$$

In practice, probabilities are estimated from a large corpus, but are “smoothed” intelligently to avoid zero probability n -grams.

Language modeling is often the art of good smoothing.

See [Goodman 1998]

$$\approx \prod_{i=1}^I p(w_i | w_{i-k}, \dots, w_{i-1})$$

The beloved n -gram language model

Slide 13 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Speech Rec = Machine Translation?

- Peter F. Brown
- Stephen A. Della Pietra
- Vincent J. Della Pietra
- Robert Mercer
- The Mathematics of Statistical Machine Translation: Parameter Estimation*
- Computational Linguistics 19 (2), June 1993

- Probably the most important paper in NLP in the last 20 years

“Brown 93”

Slide 14 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Centauri/Arcturan [Knight 97]

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

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

Slide 15 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Centauri/Arcturan [Knight 97]

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

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

Slide 16 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Centauri/Arcturan [Knight 97]

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

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

Slide 17 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Centauri/Arcturan [Knight 97]

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

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

Slide 18 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Centauri/Arcturan [Knight 97]

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

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

cognate?

Slide 25 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Centauri/Arcturan [Knight 97]

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

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

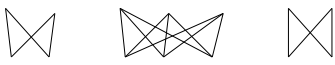
zero
fertility

Slide 26 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Unsupervised EM Training

... la maison la maison bleue la fleur ...



... the house the blue house the flower ...


All P(french-word | english-word) equally likely

Slide 27 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Unsupervised EM Training

... la maison la maison bleue la fleur ...



... the house the blue house the flower ...


“la” and “the” observed to co-occur frequently,
so $P(\text{la} | \text{the})$ is increased.

Slide 28 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Unsupervised EM Training

... la maison la maison bleue la fleur ...



... the house the blue house the flower ...

“maison” co-occurs with both “the” and “house”, but
 $P(\text{maison} | \text{house})$ can be raised without limit, to 1.0,
while $P(\text{maison} | \text{the})$ is limited because of “la”

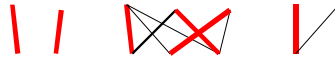
(pigeonhole principle)

Slide 29 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Unsupervised EM Training

... la maison la maison bleue la fleur ...



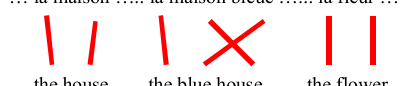
... the house the blue house the flower ...

settling down after another iteration

Slide 30 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Unsupervised EM Training

... la maison la maison bleue la fleur ...


 ... the house the blue house the flower ...

Inherent hidden structure revealed by EM training!

- "A Statistical MT Tutorial Workbook" (Knight, 1999). Promises free beer.
- "The Mathematics of Statistical Machine Translation" (Brown et al, 1993)
- Software: GIZA++

Slide 31 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

The IBM Model [Brown et al., 1993]

Mary did not slap the green witch
 Mary not slap slap slap the green witch $n(3|slap)$
 Mary not slap slap slap NULL the green witch $P NULL$
 Maria no daba una botefada a la verde bruja $t(la|the)$
 Maria no daba una botefada a la bruja verde $d(j|i)$

Use the EM algorithm for training the parameters

Slide 32 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Decoding for Machine Translation

English Source → Translation Model → French Output

$p(e)$ $p(f|e)$ $p(e|f) \propto p(e) * p(f|e)$

Decoding: $\hat{e} = \underset{e}{\operatorname{argmax}} p(e) p(f|e)$

Problem in NP-hard; use search:

- Greedy Search
- Beam Search
- Integer Programming
- A* Search

Slide 33 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Progress in Statistical MT

slide from C. Wayne, DARPA

2002	2003
<p>insistent Wednesday may recurred her trips to Libya tomorrow for flying</p> <p>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 .</p> <p>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 " .</p>	<p>Egyptair Has Tomorrow to Resume Its Flights to Libya</p> <p>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.</p> <p>" 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 " .</p>

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Automatic Evaluation of Translation

Reference 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 Tri-gram match

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. Bi-gram matches

"Bleu" metric

Slide 35 CS 5300: MT

UNIVERSITY OF UTAH
Hal Daumé III (hal@cs.utah.edu)

Minimum Error Rate Training for MT

- Desire MT system with high BLEU/??? scores [Och, ACL03]
- Algorithm:
 - Build MT system based on generative parameters
 - Decode development corpus to get n-best lists (~10k best)
 - Optimize parameters to get high BLEU scores on n-best lists
 - Repeat until converged

Slide 36 CS 5300: MT

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

Phrase-Based Translation

[Koehn, Och and Marcu, NAACL03]

Slide 37 CS 5300: MT

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

Training Phrase-Based MT Systems

[Koehn, Och and Marcu, NAACL03]

Slide 38 CS 5300: MT

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

Decoding Phrase-Based MT

[Koehn, Och and Marcu, NAACL03]

Maria no daba una botefada a la bruja verde

Mary did not slap the green witch

- Each step induces a cost attributed to:
 - Language model probability: $p(\text{slap} | \text{did not})$
 - T-table probability: $p(\text{the} | a \text{ la})$ and $p(a \text{ la} | \text{the})$
 - Distortion probability: $p(\text{skip } 1)$ [for a la --> verde]
 - Length penalty
 - ...

Slide 39 CS 5300: MT

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

Hierarchical Phrase-Based MT

[Chiang, ACL05]

Slide 40 CS 5300: MT

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

Hierarchical Phrase-Based MT

[Chiang, ACL05]

Slide 41 CS 5300: MT

UNIVERSITY OF UTAH

Hal Daumé III (hal@cs.utah.edu)

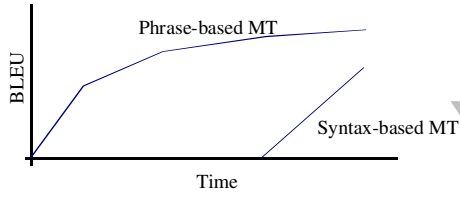
Syntax for MT

Kevin Knight, Daniel Marcu, Ignacio Thayer, Jonathan Graehl, Jon May, Steve DeNeefe

Slide 42 CS 5300: MT

Syntax for MT

- Decoding:
 - Tree-to-tree/string automata
 - CKY parsing algorithm
- Rule learning:
 - Parsed English corpus
 - Aligned data (GIZA++)
 - Extract rules and assign probabilities



Summary

- Old school translation = interlingua
 - Works well for limited domains
 - Costs a lot of money
- New school translation = statistical
 - Started out naïve
 - Becoming more linguistically motivated every year
- Translation is currently the “hot topic” in NLP
 - It looks like linguistics really is going to help, after all
 - (so long as you use it wisely in conjunction with statistics)