Introduction to Statistical Machine Translation

David Kauchak
CS159 – Spring 2011

Admin

• How did assignment 5 go?
• Project proposals?
  – I will give you feedback soon
• Start working on the projects!
• Quiz on Wednesday

Quiz #3

• text similarity
  – set overlap methods
  – vector-based methods
  – different distance metrics
  – weighting schemes: IDF and PMI
• word similarity
  – character-based
  – semantic web-based
  – dictionary-based
  – distributional/similarity-based
• misc topics:
  – stoplist
  – WordNet
  – edit distance
• information retrieval
  – general problems, evaluation, etc.
  – papers/student presentations

Language translation

Yo quiero Taco Bell
MT Systems

Where have you seen machine translation systems?

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.
People around the world stubbornly refuse to write everything in English.

Machine 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 is becoming very prevalent
Even PowerPoint has translation built into it!

The American Guam international airport and the office will receive one to call self Saudi Arabian rich merchant Radwan and so on the email which will send out, the threat can offer public place launch biochemistry attacks and so on the airport, Guam after maintenance high alert.
Beijing Youth Daily said that under the Ministry of Agriculture, the beef will be destroyed after tests.

The Beijing Youth Daily pointed out that the seized beef would be disposed of after being examined according to advice from the Ministry of Agriculture.

Pakistan President Pervez Musharraf Wins Senate Confidence Vote

Pakistani President Musharraf Won the Trust Vote in Senate and Lower House

There was not a single vote against him."

No members vote against him. *
Warren Weaver (1947)
decipherment is the analysis of documents written in ancient languages...
Warren Weaver (1947)

The non-Turkish guy next to me is even deciphering Turkish! All he needs is a statistical table of letter-pair frequencies in Turkish …

Collected mechanically from a Turkish body of text, or corpus

“... as to the problem of mechanical translation, I frankly am afraid that the [semantic] boundaries of words in different languages are too vague ... to make any quasi-mechanical translation scheme very hopeful.”

- Norbert Wiener, April 1947

“When I look at an article in Russian, I say: this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

- Warren Weaver, March 1947

MT Pyramid

interlingua

source

target

words

phrases

syntax

semantics
Data-Driven Machine Translation

Hmm, every time he sees “banco”, he either types “bank” or “bench”… but if he sees “banco de…”, he always types “bank”, never “bench”…

Man, this is so boring.

Centauri/Arcturan [Knight, 1997]
Your assignment, translate this to Arcturan:

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. orok 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 jijat quat cat.
6a. lalok sprok izok jok stok.
6b. wat dat krat quat cat.
7a. lalok farok ororok lalok sprok izok enemok.
7b. wat jij Marat vat dat vat enem.
8a. lalok brok anok plok nok.
8b. sat lat pippat mat mat.
9a. wiwok nok izok kantok ok-yurp.
9b. totat dat arrat vat hilat.
10a. lalok mok nok yorok ghirok clok.
10b. wat nnaat mat mat bat hilat.
11a. lalok rekrek hihok yorok zanzanok.
11b. wat mat mat vat zanzanat.
12a. lalok marat nok izok hikok mox.
12b. wat mat furat arrat vat jijat.

Welcome to the Chinese Room

You can teach yourself to translate Chinese using only bilingual data (without grammar books, dictionaries, any people to answer your questions…).

Centauri/Arcturan [Knight, 1997]
Your assignment, translate this to Arcturan:

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. orok 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 jijat quat cat.
6a. lalok sprok izok jok stok.
6b. wat dat krat quat cat.
7a. lalok farok ororok lalok sprok izok enemok.
7b. wat jij Marat vat dat vat enem.
8a. lalok brok anok plok nok.
8b. sat lat pippat mat mat.
9a. wiwok nok izok kantok ok-yurp.
9b. totat dat arrat vat hilat.
10a. lalok mok nok yorok ghirok clok.
10b. wat nnaat mat mat bat hilat.
11a. lalok rekrek hihok yorok zanzanok.
11b. wat mat mat vat zanzanat.
12a. lalok marat nok izok hikok mox.
12b. wat mat furat arrat vat jijat.
Your assignment, translate this to Arcturan:

1. ok-voon ororok sprok .
2. ok-drubel ok-voon anok plok sprok .
3. erok sprok izok hihok ghirok .
4. ok-voon anok drok brok jok .
5. wiwok farok izok stok .
6. wat dat latar quant cat .

7. at-voon bichat dat .
8. at-drubel an-voon poppat rat dat .
9. wat dat arrat vat jat .
10. lalok mok nok yorok ghirok clok .
11. lalok nok izok hihok ghirok .
12. lalok brok anok plok nok .

8. totat dat arrat vat hilat .
9. wat dat jat bichat wat dat vat enusat .
10. lalok nok izok hihok ok-yurp .
11. wat dat jat bichat wat dat vat enusat .
12. lalok nok izok hihok mok .

9. wat dat latar quant cat .
10. wat dat jat bichat wat dat vat enusat .
11. lalok nok izok hihok ok-yurp .
12. lalok nok izok hihok mok .

11. wat dat latar quant cat .
12. lalok nok izok hihok mok .

12. wat dat latar quant cat .
13. wat dat jat bichat wat dat vat enusat .
14. at-voon anok nok izok hihok mok .
15. lalok nok izok hihok ok-yurp .
16. at-voon anok nok izok hihok ok-yurp .
17. at-voon anok nok izok hihok ok-yurp .
18. at-voon anok nok izok hihok ok-yurp .

19. wat dat jat bichat wat dat vat enusat .
Your assignment, translate this to Arcturan:

Centauri/Arcturan [Knight, 1997]

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. latok brok anok plok nok.
6b. wat dat lat quant cat.
7a. lalok farok ororok lalok sprok izok enemok.
7b. wat jat bichat vat dat vat eneat.
8a. latok brok anok plok nok.
8b. iat lat pippat rat mat.
9a. wiwok nok izok kantok ok-yurp.
9b. totat nnat quat oloat at-yurp.
10a. lalok farok ororok lalok sprok izok enemok.
10b. wat nnat gat mat bat hilat.
11a. lalok nok izok hihok yorok zanzanok.
11b. wat nnat arrat mat zanzanat.
12a. lalok nok izok hihok yorok clok.
12b. wat nnat forat arrat vat gat.

Your assignment, translate this to Arcturan:

Centauri/Arcturan [Knight, 1997]

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. latok brok anok plok nok.
6b. wat dat lat quant cat.
7a. lalok farok ororok lalok sprok izok enemok.
7b. wat jat bichat vat dat vat eneat.
8a. latok brok anok plok nok.
8b. iat lat pippat rat mat.
9a. wiwok nok izok kantok ok-yurp.
9b. totat nnat quat oloat at-yurp.
10a. lalok farok ororok lalok sprok izok enemok.
10b. wat nnat gat mat bat hilat.
11a. lalok nok izok hihok yorok zanzanok.
11b. wat nnat arrat mat zanzanat.
12a. lalok nok izok hihok mok.
12b. wat nnat forat arrat vat gat.
Your assignment, translate this to Arcturan:

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

2a. ok-drubel ok-voon anok plok sprok.
2b. at-drubel at-voon peppat rat dat.

3a. وكوك وكوك وكوك وكوك.
3b. at-drubel at-voon pippat rrat dat.

4a. ok-voon anok druk brok jot.
4b. at-voon krat peppat sat lat.

5a. ok-voon anok druk brok jot.
5b. at-voon krat peppat sat lat.

6a. ok-voon anok druk brok jot.
6b. at-voon krat peppat sat lat.

7a. ok-voon ororok sprok izok enemok.
7b. at-voon bichat wat dat vat eneat.

8a. ok-voon ororok sprok izok enemok.
8b. at-voon bichat wat dat vat eneat.

9a. ok-voon ororok sprok izok enemok.
9b. at-voon bichat wat dat vat eneat.

10a. ok-voon ororok sprok izok enemok.
10b. at-voon bichat wat dat vat eneat.

11a. ok-voon ororok sprok izok enemok.
11b. at-voon bichat wat dat vat eneat.

12a. ok-voon ororok sprok izok enemok.
12b. at-voon bichat wat dat vat eneat.

Your assignment, put these words in order:

{jjat, arrat, mat, bat, oloat, at-yurp}

Your assignment, translate this to Arcturan:

1a. wiwok nok izok stok.
1b. to totat dat arrat vat hilat.

2a. wiwok nok izok stok.
2b. to totat dat arrat vat hilat.

3a. wiwok nok izok stok.
3b. to totat dat arrat vat hilat.

4a. wiwok nok izok stok.
4b. to totat dat arrat vat hilat.

5a. wiwok nok izok stok.
5b. to totat dat arrat vat hilat.

6a. wiwok nok izok stok.
6b. to totat dat arrat vat hilat.

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.

2a. Carlos Garcia has three associates.
2b. Carlos Garcia tiene tres asociados.

3a. his associates are not strong.
3b. sus asociados no son fuertes.

4a. Garcia has a company also.
4b. Garcia tambien tiene una empresa.

5a. his clients are angry.
5b. sus clientes estan enfadados.

6a. the associates are also angry.
6b. sus asociados tambien estan enfadados.

7a. the clients are enemies.
7b. los clientes y los asociados son enemigos.

8a. the company has three groups.
8b. la empresa tiene tres grupos.

9a. its clients are in Europe.
9b. sus grupos estan en Europa.

10a. the modern groups sell strong pharmaceuticals.
10b. los grupos modernos venden medicinas fuertes.
Data available

- Many languages
  - Europarl corpus has all European languages
    - http://www.statmt.org/europarl/
    - From a few hundred thousand sentences to a few million
  - French/English from French parliamentary proceedings
  - Lots of Chinese/English and Arabic/English from government projects/interests
    - Chinese-English: 440 million words (15-20 million sentence pairs)
    - Arabic-English: 790 million words (30-40 million sentence pairs)
  - Smaller corpora in many, many other languages

- Lots of monolingual data available in many languages
- Even less data with multiple translations available
- Available in limited domains
  - most data is either news or government proceedings
  - some other domains recently, like blogs

Statistical MT Overview

Bilingual data

monolingual data

model

training

learned parameters

Translation

Foreign sentence

Find the best translation given the foreign sentence and the model

English sentence

Statistical MT

- We will model the translation process probabilistically
- Given a foreign sentence to translate, for any possible English sentence, we want to know what the probability that sentence is a translation of the foreign sentence
- If we can find the most probable English sentence, we’re done

\[ p(\text{english sentence} \mid \text{foreign sentence}) \]

Noisy channel model

some message is sent

along the way the message gets messed up

What was originally sent?

We have the mutated message, but would like to recover the original
Noisy channel model

\[ p(s | r) = \frac{p(f | e)p(e)}{p(f)} \]

Bayes’ rule

- \( p(f) \): probability of the foreign sentence
- \( p(e) \): language model: what are likely English word sequences?
- \( p(f | e) \): translation model: how does the translation process happen? probability of the translated English sentence given the foreign sentence
Noisy channel model

\[ p(e \mid f) \propto p(f \mid e) p(e) \]

- Translation model
- Language model

How do foreign sentences get translated to English sentences?

What do English sentences look like?

Translation model

- The models define probabilities over inputs
  \[ p(f \mid e) \]

Morgen fliege ich nach Kanada zur Konferenz

Tomorrow I will fly to the conference in Canada

What is the probability that the English sentence is a translation of the foreign sentence?

Translation model

- The models define probabilities over inputs
  \[ p(f \mid e) \]

What is the probability of a foreign word being translated as a particular English word?

What is the probability of a foreign foreign phrase being translated as a particular English phrase?

What is the probability of a word/phrase changing ordering?

What is the probability of a foreign word/phrase disappearing?

What is the probability of an English word/phrase appearing?
Translation model

- The models define probabilities over inputs
  \[ p(f | e) \]

\[ p(\text{Morgen fliege ich nach Kanada zur Konferenz} | \text{Tomorrow I will fly to the conference in Canada}) = 0.1 \]

\[ p(\text{Morgen fliege ich nach Kanada zur Konferenz} | \text{I like peanut butter and jelly}) = 0.0001 \]

Language model

- The models define probabilities over inputs
  \[ p(e) \]

\[ \text{Tomorrow I will fly to the conference in Canada} \]

What is a probability distribution?

- A probability distribution defines the probability over a space of possible inputs
- For the language model, what is the space of possible inputs?
  - A language model describes the probability over ALL possible combinations of English words
- For the translation model, what is the space of possible inputs?
  - ALL possible combinations of foreign words with ALL possible combinations of English words

One way to think about it…

Spanish (foreign) → Translation model → Broken English → Language model → English

Que hambre tengo yo → What hunger have I,
Hungry I am so,
I am so hungry,
Have I that hunger … → I am so hungry
Translation

\[ p(e \mid f) \propto p(f \mid e)p(e) \]

- Let’s assume we have a translation model and a language model
- Given a foreign sentence, what question do we want to ask to translate that sentence into English?
  \[ \arg\max_e p(e \mid f) \propto p(f \mid e)p(e) \]

**Statistical MT Overview**

Basic Model, Revisited

\[
\begin{align*}
\text{argmax } P(e \mid f) &= e \\
\text{argmax } P(e) \times P(f \mid e) / P(f) &= e \\
\text{argmax } P(e) \times P(f \mid e) &= e
\end{align*}
\]

Basic Model, Revisited

\[
\begin{align*}
\text{argmax } P(e \mid f) &= e \\
\text{argmax } P(e) \times P(f \mid e) / P(f) &= e \\
\text{argmax } P(e)^2 \times P(f \mid e) &= e
\end{align*}
\]... works better!
Basic Model, Revisited

argmax \ P(e | f) =
\ e

argmax \ P(e) x P(f | e) / P(f)
\ e

argmax \ P(e) x P(f | e) x length(e)^{1.1}
\ e

Rewards longer hypotheses, since these are unfairly punished by P(e).

Basic Model, Revisited

argmax \ P(e)^{2.4} x P(f | e) x length(e)^{1.1} x KS^{3.7} ...
\ e

Lots of knowledge sources vote on any given hypothesis. “Knowledge source” = “feature function” = “score component”. A feature function simply scores a hypothesis with a real value. (May be binary, as in “e has a verb”).

Problems for Statistical MT

• Preprocessing
  – How do we get aligned bilingual text?
  – Tokenization (document, sentence, word)
• Language modeling
  – Given an English string e, assigns P(e) by formula
• Translation modeling
  – Given a pair of strings <f,e>, assigns P(f | e) by formula
• Decoding
  – Given a language model, a translation model, and a new sentence f ... find translation e maximizing P(e) * P(f | e)
• Parameter optimization
  – Given a model with multiple feature functions, how are they related? What are the optimal parameters?
• Evaluation
  – How well is a system doing? How can we compare two systems?
Problems for Statistical MT

• Preprocessing
• Language modeling
• Translation modeling
• Decoding
• Parameter optimization
• Evaluation

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)

If you don’t get the characters right…

ISO-8859-2 (Latin2)

ISO-8859-6 (Arabic)
Chinese?

- GB Code
- GBK Code
- Big 5 Code
- CNS-11643-1992
- ...

Document Alignment

- Input:
  - Big bag of files obtained from somewhere, believed to contain pairs of files that are translations of each other.
- Output:
  - List of pairs of files that are actually translations.

Document Alignment

- Input:
  - Big bag of files obtained from somewhere, believed to contain pairs of files that are translations of each other.
- Output:
  - List of pairs of files that are actually translations.

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.
2. Ha pescado muchas veces.
3. Su mujer habla con él.
4. 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.

Tokenization (or Segmentation)

• English
  – Input (some byte stream):
    "There," said Bob.
  – Output (7 "tokens" or "words"):
    "There, " said Bob.

• Chinese
  – Input (byte stream):
    美国洛杉矶机场及其办公室发布的
    一名台湾籍在洛杉矶机场的男性
    没有外包的电子邮件。
  – Output:
    美国洛杉矶机场及其办公室发布的
    一名台湾籍在洛杉矶机场的男性
    没有外包的电子邮件。
Problems for Statistical MT

• Preprocessing
• Language modeling
• Translation modeling
• Decoding
• Parameter optimization
• Evaluation

Language Modeling

• Most common: n-gram language models
• More data the better (Google n-grams)
• Domain is important

Problems for Statistical MT

• Preprocessing
• Language modeling
• Translation modeling
• Decoding
• Parameter optimization
• Evaluation
Translation Model
Learn How to Translate from Data

Direct Estimation:

Mary did not slap the green witch

not enough data for this (most input sentences unseen)

Maria no dió una botefada a la bruja verde

Generative Model
Break up process into smaller steps:

Mary did not slap the green witch

sufficient statistics for smaller steps

Maria no dió una botefada a la bruja verde

What kind of Translation Model?
May use syntactic and semantic representations:

Word-level models
Phrasal models
Syntactic models
Semantic models

Mary did not slap the green witch

Maria no dió una botefada a la bruja verde

The Classic Translation Model
Word Substitution/Permutation [IBM Model 3, Brown et al., 1993]

Generative story:

Mary did not slap the green witch

r(3|slap)
P-Null
t(la|the)
d(j|i)

Maria no dió una botefada a la bruja verde

Probabilities can be learned from raw bilingual text.
Phrase-Based Statistical MT

- Foreign input segmented into phrases
  - "phrase" is any sequence of words
- Each phrase is probabilistically translated into English
  - $P(\text{to the conference} \mid \text{zur Konferenz})$
  - $P(\text{into the meeting} \mid \text{zur Konferenz})$
- Phrases are probabilistically re-ordered

Advantages of Phrase-Based

- Many-to-many mappings can handle non-compositional phrases
- Easy to understand
- Local context is very useful for disambiguating
  - "Interest rate" $\rightarrow$ …
  - "Interest in" $\rightarrow$ …
- The more data, the longer the learned phrases
  - Sometimes whole sentences

Syntax

These 7 people include astronauts coming from France and Russia.
Problems for Statistical MT

- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation

Decoding

- Of all conceivable English word strings, find the one maximizing $P(e) \times P(f \mid e)$
- Decoding is an NP-complete problem (for many translation models – (Knight, 1999)
- Several decoding strategies are often available

Search

$\arg\max \ p(f \mid e)p(e)$

Search

$\arg\max \ p(f \mid e)p(e)$

partial trans.

$\cdots$
Evaluation
Parameter optimization
Decoding
Translation modeling
Language modeling
Preprocessing

Problems for Statistical MT
• Preprocessing
• Language modeling
• Translation modeling
• Decoding
• Parameter optimization
• Evaluation

The Problem: Learn Lambdas

\[
p(e | f) = \frac{p(f | e)p(e)}{p(f)} = \frac{p(f | e)p(e)}{\sum p(f | e') \lambda_p(e')}
\]

\[
- \frac{p(f | e) \exp \lambda_p(f | e') p(e) \exp \lambda_p(f | e') \exp \lambda_p(f | e') \exp \lambda_p(f | e') \exp \lambda_p(f | e') \exp \lambda_p(f | e') \exp \lambda_p(f | e')}{\sum p(f | e') \exp \lambda_p(f | e') \exp \lambda_p(f | e') \exp \lambda_p(f | e') \exp \lambda_p(f | e') \exp \lambda_p(f | e') \exp \lambda_p(f | e')}
\]

Given a data set with foreign/English sentences, find the \( \lambda \)'s that:
• maximize the likelihood of the data
• maximize an evaluation criterion
Problems for Statistical MT

- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation

MT Evaluation

- Source only
- Manual:
  - SSER (subjective sentence error rate)
  - Correct/Incorrect
  - Error categorization
- Objective usage testing
- Automatic:
  - WER (word error rate)
  - BLEU (Bilingual Evaluation Understudy)
  - NIST
  - Named-Entity

Reference Evaluation

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.

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

- 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 a 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)
The U.S. island of Guam is maintaining a high state of alert after public places as 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. Guam would be biochemistry air raid to Guam, which threats will be able to start the biochemical attack. [?] The threat will be able to start the biochemical attack. [?] Highly threats after the maintenance.

---

**BLEU Evaluation Metric**

**Reference (human) translation:**
- Machine 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 threat will be able to start the biochemical attack.

**BLEU formula**
- Generally N=4
- \( p_i^{N} \) uniform weights
- \( BLEU = N \prod_{i=1}^{N} p_i^{N} \)
- \( BP \) = brevity penalty
- \( p_i \) = i-gram precision
- \( = \frac{1}{N} \) (uniform weights)

**Multiple Reference Translations**

**Reference translation 1:**
- Reference translation 2:
- Reference translation 3:
- Reference translation 4:

---

**Available Resources**

- **Bilingual corpora**
  - 100+ million words of Chinese/English and Arabic/English, LDC (www.ldc.upenn.edu)
  - Lots of French/English, Spanish/English, LDC
  - European Parliament (sentence-aligned), 11 languages, Philipp Koehn, ISI
  - (www.isi.edu/~koehn/publications/typocept)
  - (www.isi.edu/~koehn/publications/typocept)

- **Sentence alignment**
  - Dan Melamed, NYU (www.cs.nyu.edu/~melamed/GMA/docs/README.htm)
  - Xiaoyi Ma, LDC (Champion)

- **Word alignment**
  - GIZA++, JHU Workshop ’99 (www.clsp.jhu.edu/ws99/projects/mt)
  - GIZA++, RWTH Aachen (www.iwi.informatik.rwth-aachen.de/Software/GIZA++
  - Lexical translation training corpus (50 French/English sentence pairs), RWTH Aachen
  - Shared task, NAACL-HLT’03 workshop

- **Decoding**
  - ISI ReWrite Model 4 decoder (www.isi.edu/~knight/rewrite-decoder)
  - ISI Phrasal phrase-based decoder

- **Statistical MT Tutorial Workbooks**
  - ISI (www.isi.edu/~knight)
  - Annual common-data evaluation, NIST (www.nist.gov/speech/tests/mt/index.htm)

---

**Some Papers Referenced on Slides**

- **ACL**
  - [Och, Titman, & Ney, 1999]
  - [Och & Ney, 2000]
  - [Germann et al., 2001]
  - [Yamada & Knight, 2001, 2002]
  - [Papineni et al., 2002]
  - [Al-Onaizan & Knight, 1998]
  - [Collins, 1997]
  - [Koehn & Knight, 2003]
  - [Al-Onaizan & Knight, 2002]
  - [Och & Ney, 2002]
  - [Och, 2003]
  - [Koehn et al., 2003]

- **EMNLP**
  - [Marcu & Wong, 2002]
  - [Fox, 2002]
  - [Och, 2003]
  - [Och, 2003]
  - [Munteanu & Marcu, 2002]
  - [Koehn et al., 2003]

- **Al Magazine**
  - [Knight, 1997]
  - [MIT Tutorial Workbook]

- **MT Summit**
  - [Charniak, Knight, Yamada, 2003]

- **NAACL**
  - [Koehn, Marcu, Och, 2003]
  - [Germann, 2003]
  - [Och et al., 2002]
  - [Och, 2003]

- **AMTA**
  - [Soricut et al., 2002]
  - [Al-Onaizan & Knight, 1998]

- **EACL**
  - [Cremieux et al., 2003]

- **Computational Linguistics**
  - [Brown et al., 1993]
  - [Knight, 1999]
  - [Wu, 1997]

- **AAAI**
  - [Koehn & Knight, 2000]

- **IWCL**
  - [Och, 2002]

- **MT Summit**
  - [Charniak, Knight, Yamada, 2003]

- **NAACL**
  - [Koehn, Marcu, Och, 2003]
  - [Germann, 2003]
  - [Knight, 2004]