Introduction to Statistical Machine Translation

David Kauchak
CS159 – Fall 2014

Some slides adapted from

Philipp Koehn
School of Informatics
University of Edinburgh

Kevin Knight
USC/Information Sciences Institute
USC/Computer Science Department

Dan Klein
Computer Science Department
UC Berkeley

Admin

Assignment 5

Normal office hours on Friday

Language translation

MT Systems

Where have you seen machine translation systems?

Machine Translation

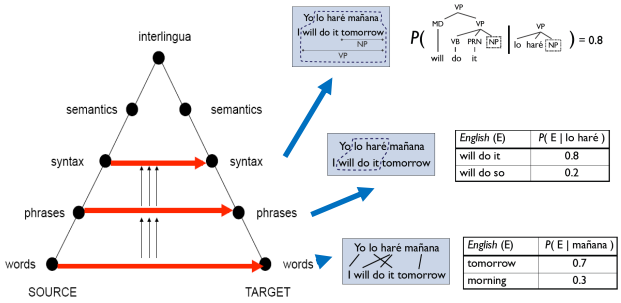
美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

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 .

A good test for natural language processing.

Requires capabilities in both interpretation and generation.

Levels of Transfer



World-Level MT: Examples

la politique de la haine .	*(Foreign Original)*
politics of hate .	(Reference Translation)
the policy of the hatred .	(IBM4+N-grams+Stack)
nous avons signé le protocole .	*(Foreign Original)*
we did sign the memorandum of agreement .	(Reference Translation)
we have signed the protocol .	(IBM4+N-grams+Stack)
où était le plan solide ?	*(Foreign Original)*
but where was the solid plan ?	(Reference Translation)
where was the economic base ?	(IBM4+N-grams+Stack)

Phrasal / Syntactic MT: Examples

Le président américain Barack Obama doit annoncer lundi de nouvelles mesures en faveur des constructeurs automobile. General motors et Chrysler avaient déjà bénéficié fin 2008 d'un prêt d'urgence cumulé de 17,4 milliards de dollars, et ont soumis en février au Trésor un plan de restructuration basé sur un total de 22 milliards de dollars d'aides publiques supplémentaires.

Interrogé sur la chaîne CBS dimanche, le président a toutefois clairement précisé que le gouvernement ne preterait pas d'argent sans de fortes contreparties. *"Il faudra faire des sacrifices à tous les niveaux"*, a-t-il prévenu. *"Tout le monde devra se réunir autour de la table et se mettre d'accord sur une restructuration en profondeur"*.

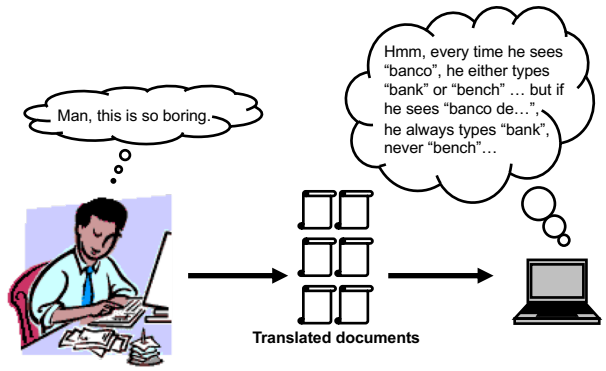
General Motors et Chrysler sont engagés dans des négociations avec le principal syndicat de l'automobile. Les constructeurs souhaitent diminuer leurs cotisations aux caisses de retraites, et accorder en échange des actions aux syndicats. Ils souhaiteraient également négocier des baisses des salaires.

U.S. President Barack Obama to announce Monday new measures to help automakers. General Motors and Chrysler had already received late in 2008 a cumulative emergency loan of 17.4 billion dollars, and submitted to the Treasury in February in a restructuring plan based on a total of 22 billion dollars in additional aid .

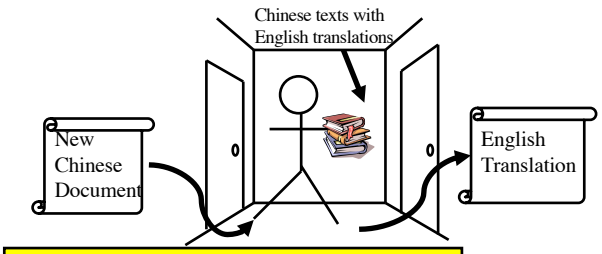
Interviewed on CBS Sunday, the president has clearly stated that the government does not lend money without strong counterparts. *"We must make sacrifices at all levels,"* he warned. *"Everyone should gather around the table and agree on a profound restructuring. "*

General Motors and Chrysler are engaged in negotiations with the major union of the car. Manufacturers wishing to reduce their contributions to pension funds, and give in exchange for the shares to trade unions. They would also negotiate lower wages.

Data-Driven Machine Translation



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: 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 .	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 .	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 .	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 .	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 .	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 .	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 .	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 .	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 .	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 . process of elimination
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 .	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 available

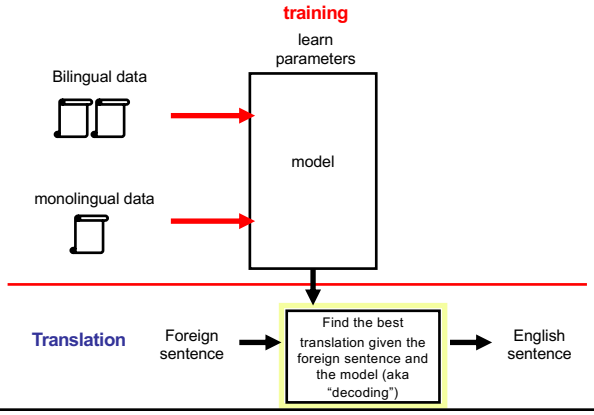
Many languages
- Europarl corpus has all European languages
  - http://www.statmt.org/europarl/
  - From a few hundred thousand sentences to a few million
- French/English from French parliamentary proceedings
- Lots of Chinese/English and Arabic/English from government projects/interests
  - Chinese-English: Hundreds of millions if sentence pairs)
  - Arabic-English: ~One hundred 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



Statistical MT

We will model the translation process probabilistically

Given a foreign sentence to translate, for any possible English sentence, we want to know the probability that the sentence is a translation of the foreign sentence

If we can find the most probable English sentence, we're done

p(english sentence | foreign sentence)

Translation

Probabilistic model: $p(e \mid f)$ p(English | Foreign)

What is the translation problem then?

$$translation(f) = \arg_e \max p(e \mid f)$$

Noisy channel model

$$p(e \mid f) = \frac{p(f \mid e)p(e)}{p(f)}$$ Bayes' rule

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

Noisy channel model

$p(e \mid f) = p(f \mid e)p(e)$ Bayes' rule

$p(f)$ probability of the foreign sentence — why?

$p(e)$ language model: what are likely English word sequences?

$p(f \mid e)$ translation model: how does the translation process happen? probability of the translated English sentence given the foreign sentence

Noisy channel model

$p(e \mid f) = p(f \mid e)p(e)$ Bayes' rule

$p(f)$ probability of the foreign sentence — why?

$$translation(f) = \arg_e \max \frac{p(f \mid e)p(e)}{p(f)} = \arg_e \max p(f \mid e)p(e)$$

this is a constant for any given f

Noisy channel model

model $p(e \mid f) \propto p(f \mid e)p(e)$

translation model: how do English sentences get translated to foreign?

language model: 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

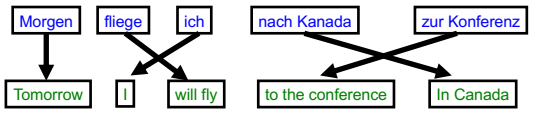
Tomorrow I will fly to the conference in Canada

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 a English word/phrase appearing?

Translation model

The models define probabilities over inputs

$$p(f \mid e)$$

p(Morgen fliege ich nach Kanada zur Konferenz | Tomorrow I will fly to the conference in Canada) = 0.1

p(Morgen fliege ich nach Kanada zur Konferenz | I like peanut butter and jelly) = 0.0001

Language model

The models define probabilities over inputs

$$p(e)$$

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

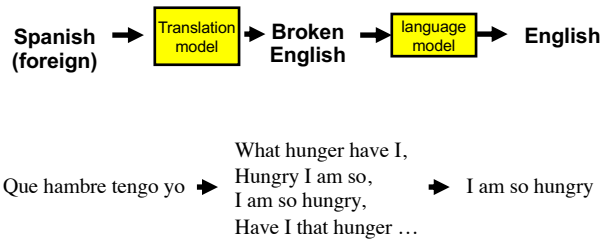
For the language model, what is the space of possible inputs?

- A language model describes the probability over **ALL** possible combinations of English words

For the translation model, what is the space of possible inputs?

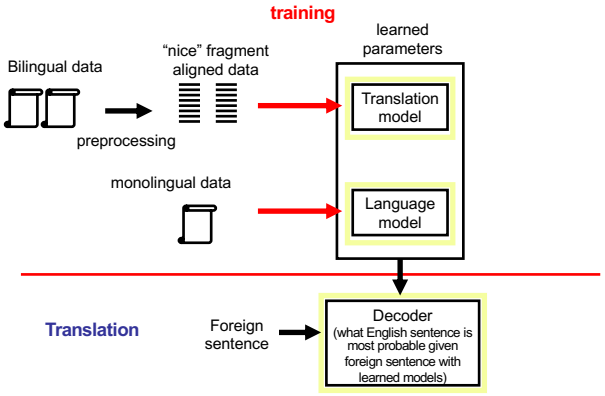
- **ALL** possible combinations of foreign words with **ALL** possible combinations of English words

One way to think about it…



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

Statistical MT Overview



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)
e

Basic Model, Revisited

argmax P(e | f) =
e

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

argmax $P(e)^{2.4}$ x P(f | e) … works better!
e

Basic Model, Revisited

$$\underset{e}{\text{argmax}}\ P(e \mid f) =$$

$$\underset{e}{\text{argmax}}\ P(e) \times P(f \mid e) / P(f)$$

$$\underset{e}{\text{argmax}}\ P(e)^{2.4} \times P(f \mid e) \times \text{length}(e)^{1.1}$$

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

Basic Model, Revisited

$$\underset{e}{\text{argmax}}\ P(e)^{2.4} \times P(f \mid e) \times \text{length}(e)^{1.1} \times KS^{3.7} \ldots$$

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
- Segmentation (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?

Translation Model

Want: probabilistic model gives us how likely one sentence is to be a translation of another, i.e *p(foreign | english)*

Mary did not slap the green witch

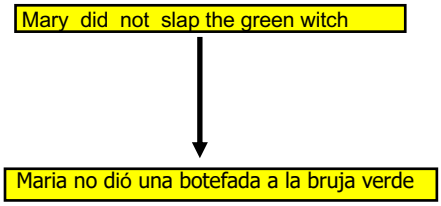
↓

Maria no dió una botefada a la bruja verde

Can we just model this directly, i.e. p(foreign | english) ?
How would we estimate these probabilities, e.g.
p("Maria …" | "Mary …")?

Translation Model

Want: probabilistic model gives us how likely one sentence is to be a translation of another, i.e *p(foreign | english)*

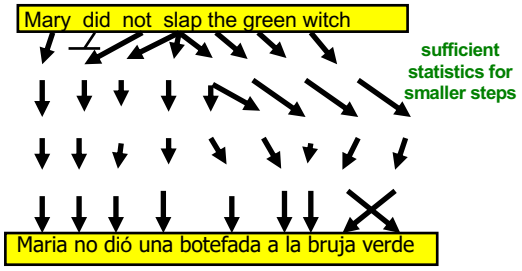


$$p(\text{“Maria…”} \mid \text{“Mary…”}) = \frac{\text{count(“Mary…” aligned-to “Maria…”)}}{\text{count(“Mary…”)}}$$

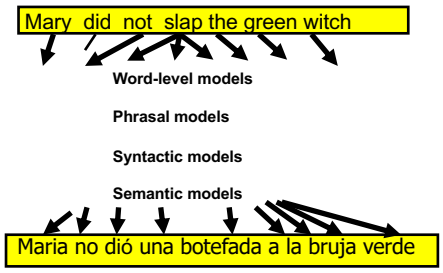
Not enough data for most sentences!

Translation Model

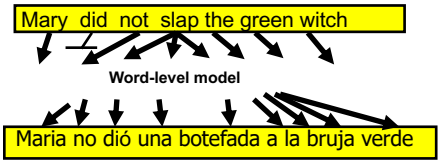
Key: break up process into smaller steps



What kind of Translation Model?



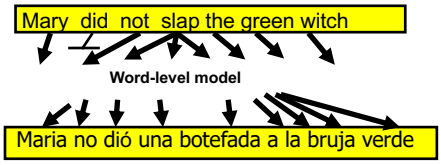
IBM Word-level models



Generative story: description of how the translation happens

1. Each English word gets translated as 0 or more Foreign words
2. Some additional foreign words get inserted
3. Foreign words then get shuffled

IBM Word-level models



Each foreign word is *aligned* to exactly one English word.

Key idea: decompose *p(foreign | english)* into word translation probabilities of the form *p(foreign_word | english_word)*

IBM described 5 different levels of models with increasing complexity (and decreasing independence assumptions)

Some notation

$E = e_1 e_2 ... e_{|E|}$ English sentence with length |E|

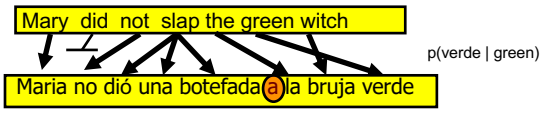
$F = f_1 f_2 ... f_{|F|}$ Foreign sentence with length |F|

Mary did not slap the green witch
e_1 e_2 e_3 e_4 e_5 e_6 e_7

f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_9
Maria no dió una botefada a la bruja verde

Translation model: $p(F \mid E) = p(f_1 f_2 ... f_{|F|} \mid e_1 e_2 ... e_{|E|})$

Word models: IBM Model 1

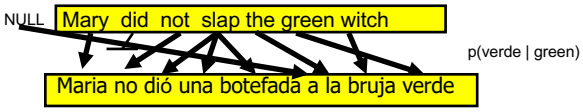


Each foreign word is aligned to exactly one English word

This is the **ONLY** thing we model!

Does the model handle foreign words that are not aligned, e.g. “a”?

Word models: IBM Model 1



Each foreign word is aligned to exactly one English word

This is the **ONLY** thing we model!

Include a “NULL” English word and align to this to account for deletion

Word models: IBM Model 1

generative story -> probabilistic model

– Key idea: introduce "hidden variables" to model the word alignment

$$p(f_1 f_2 ... f_{|F|} \mid e_1 e_2 ... e_{|E|})$$

$$p(f_1 f_2 ... f_{|F|}, a_1 a_2 ... a_{|F|} \mid e_1 e_2 ... e_{|E|})$$

- one variable for each foreign word
- a_i corresponds to the ith foreign word
- each a_i can take a value 0…|E|

Alignment variables

Mary did not slap the green witch

e_0 e_1 e_2 e_3 e_4 e_5 e_6 e_7

f_1 f_2 f_3 f_4 f_5 f_6 f_7 f_8 f_9

Maria no dió una botefada a la bruja verde

a_1	1
a_2	3
a_3	4
a_4	4
a_5	4
a_6	0
a_7	5
a_8	7
a_9	6

Alignment variables

And the program has been implemented

e_0 e_1 e_2 e_3 e_4 e_5 e_6

Alignment?

f_1 f_2 f_3 f_4 f_5 f_6 f_7

Le programme a ete mis en application

Alignment variables

And the program has been implemented

e_0 e_1 e_2 e_3 e_4 e_5 e_6

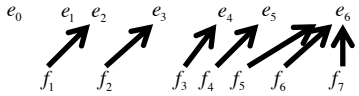
f_1 f_2 f_3 f_4 f_5 f_6 f_7

Le programme a ete mis en application

a_1	?
a_2	?
a_3	?
a_4	?
a_5	?
a_6	?
a_7	?

Alignment variables

And the program has been implemented



Le programme a ete mis en application

a_1	2
a_2	3
a_3	4
a_4	5
a_5	6
a_6	6
a_7	6

Probabilistic model

$$p(f_1 f_2 ... f_{|F|} \mid e_1 e_2 ... e_{|E|}) \stackrel{?}{=} p(f_1 f_2 ... f_{|F|}, a_1 a_2 ... a_{|F|} \mid e_1 e_2 ... e_{|E|})$$

NO!

$$p(f_1 f_2 ... f_{|F|}, a_1 a_2 ... a_{|F|} \mid e_1 e_2 ... e_{|E|}) \Rightarrow p(f_1 f_2 ... f_{|F|} \mid e_1 e_2 ... e_{|E|})$$

How do we get rid of variables?

Joint distribution

NLPPass, EngPass	P(NLPPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

What is P(ENGPass)?

Joint distribution

NLPPass, EngPass	P(NLPPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

0.92

How did you figure that out?

Joint distribution

$$P(x) = \sum_{y \in Y} p(x, y)$$

Called “marginalization”, aka summing over a variable

NLPPass, EngPass	P(NLPPass, EngPass)
true, true	.88
true, false	.01
false, true	.04
false, false	.07

Probabilistic model

$$p(f_1 f_2 \ldots f_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \sum_{a_1} \sum_{a_2} \ldots \sum_{a_{|F|}} p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|})$$

Sum over all possible values, i.e. marginalize out the alignment variables

Independence assumptions

IBM Model 1:

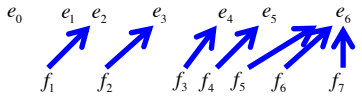
$$p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

What independence assumptions are we making?

What information is lost?

$$p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

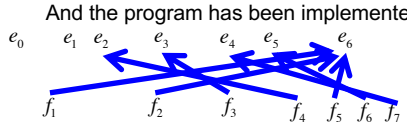
And the program has been implemented



Le programme a ete mis en application

Are the probabilities any different under model 1?

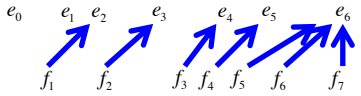
And the program has been implemented



application en programme Le mis ete a

$$p(f_1 f_2 ... f_{|F|}, a_1 a_2 ... a_{|F|} \mid e_1 e_2 ... e_{|E|}) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

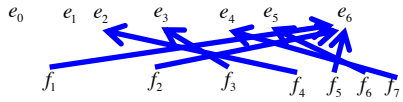
And the program has been implemented



Le programme a ete mis en application

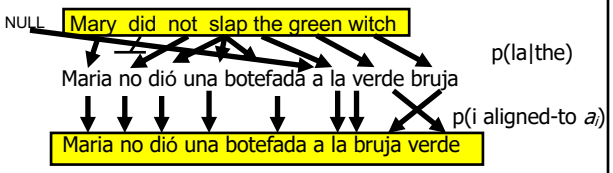
No. Model 1 ignores word order!

And the program has been implemented



application en programme Le mis ete a

IBM Model 2

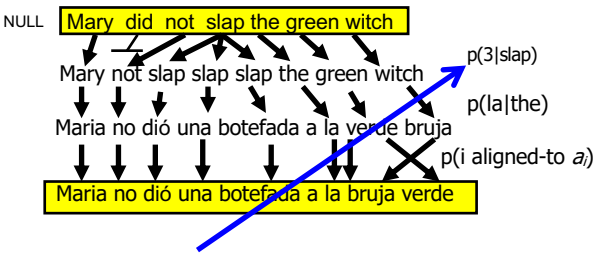


$$p(f_1 f_2 ... f_{|F|}, a_1 a_2 ... a_{|F|} \mid e_1 e_2 ... e_{|E|}) = \prod_{i=1}^{|F|} p(i \text{ aligned-to } a_i) p(f_i \mid e_{a_i})$$

Models word movement by position, e.g.

- Words don't tend to move too much
- Words at the beginning move less than words at the end

IBM Model 3



Incorporates "fertility": how likely a particular English word is to produce multiple foreign words

Word-level models

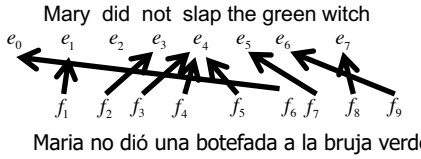
Problems/concerns?

- Multiple English words for one French word
  - IBM models can do one-to-many (fertility) but not many-to-one
- Phrasal Translation
  - "real estate", "note that", "interest in"
- Syntactic Transformations
  - Verb at the beginning in Arabic
  - Translation model penalizes any proposed re-ordering
  - Language model not strong enough to force the verb to move to the right place

Benefits of word-level model

Rarely used in practice for modern MT systems

Why talk about them?



Two key side effects of training a word-level model:

- Word-level alignment
- p(f | e): translation dictionary

Training a word-level model

$$p(f_1 f_2 \ldots f_{|F|}, a_1 a_2 \ldots a_{|F|} \mid e_1 e_2 \ldots e_{|E|}) = \prod_{i=1}^{|F|} p(f_i \mid e_{a_i})$$

Where do these come from?

Have to learn them!

The old man is happy. He has fished many times.	———	El viejo está feliz porque ha pescado muchos veces.
His wife talks to him.	———	Su mujer habla con él.
The sharks await.	———	Los tiburones esperan.
…		…