Finite-State Translation and Discriminative Training

Trevor Cohn
Based on slides by Adam Lopez
Quick Recap

training data
(parallel text)

learner

model

decoder

However, the sky remained clear under the strong north wind.

Thursday, 24 November 2011
However, the sky remained clear under the strong north wind.
Quick Recap

For our purposes, a model will be a probability distribution over English sentences given a Chinese sentence.
Quick Recap

\[ p(\text{English}|\text{Chinese}) \sim \]

\[ p(\text{English}) \times p(\text{Chinese}|\text{English}) \]

language model

translation model
Quick Recap
Quick Recap

- Language models
Quick Recap

- Language models
  - Each word is dependent only on $n-1$ preceding words (called an $n$-gram language model).
Quick Recap

- Language models
  - Each word is dependent only on $n-1$ preceding words (called an $n$-gram language model).
  - Requires only monolingual text, available in vast quantities -- more data seems to help.
Quick Recap

● Language models

● Each word is dependent only on \( n-1 \) preceding words (called an \( n \)-gram language model).

● Requires only monolingual text, available in vast quantities -- more data seems to help.

● Translation models
Quick Recap

- **Language models**
  - Each word is dependent only on \( n-1 \) preceding words (called an \( n \)-gram language model).
  - Requires only monolingual text, available in vast quantities -- more data seems to help.

- **Translation models**
  - Seen word-based translation models
Quick Recap

- Language models
  - Each word is dependent only on \( n-1 \) preceding words (called an \( n \)-gram language model).
  - Requires only monolingual text, available in vast quantities -- more data seems to help.

- Translation models
  - Seen word-based translation models
  - More to come...
The IBM Models
The IBM Models

• Fertility probabilities.
The IBM Models

- Fertility probabilities.
- Word translation probabilities.
The IBM Models

- Fertility probabilities.
- Word translation probabilities.
- Distortion probabilities.
The IBM Models

- Fertility probabilities.
- Word translation probabilities.
- Distortion probabilities.
The IBM Models

- Fertility probabilities.
- Word translation probabilities.
- Distortion probabilities.

Some problems:
- Weak reordering model -- output is not fluent.
- Many decisions -- many things can go wrong.
Although north wind howls, but sky still very clear. However, the sky remained clear under the strong north wind.
IBM Model 4 Again

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Can we do better?
Garcia and associates.

Carlos García has three associates.

His associates are not strong.

Garcia has a company also.

Its clients are angry.

The associates are also angry.

The company has three groups.

Its groups are in Europe.

The modern groups sell strong pharmaceuticals.

The groups do not sell zanzanine.

The small groups are not modern.

La empresa tiene enemigos fuertes en Europa.
Garcia and associates.

Carlos García has three associates.

His associates are not strong.

Garcia has a company also.

It's clients are angry.

The associates are also angry.

The company has three groups.

Its groups are in Europe.

The modern groups sell strong pharmaceuticals.

The groups do not sell zanzanine.

The small groups are not modern.

La empresa tiene enemigos fuertes en Europa.
Garcia and associates.
The clients and the associates are enemies.

Carlos Garcia has three associates.

His associates are not strong.

Garcia has a company also.

Garcia y asociados.

Los clientes y los asociados son enemigos.

Carlos Garcia tiene tres asociados.

Sus asociados no son fuertes.

Garcia tambien tiene una empresa.

Los grupos modernos venden medicinas fuertes.

The modern groups sell strong pharmaceuticals.

Sus clientes estan enfadados.

Los clientes y los asociados son enemigos.

The groups do not sell zanzanine.

Los grupos no venden zanzanina.

The small groups are not modern.

Los grupos pequenos no son modernos.

The company has three groups.

La empresa tiene tres grupos.

Sus grupos estan en Europa.

Its groups are in Europe.
Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Thursday, 24 November 2011
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

虽然北风呼啸，但天空依然十分清澈。

虽然 ▶️ However
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

虽然北风呼啸，但天空依然十分清澈。

However, the sky remained clear under the strong north wind.
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.
Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。

However, the sky remained clear under the strong north wind.

虽然北风呼啸，但天空依然十分清澈

Thursday, 24 November 2011
Phrase-Translation Distribution

- Collect all the extracted phrase pairs
- Normalise frequencies to get translation scores

$$\phi(f|e) = \frac{c(f, e)}{c(\cdot, e)}$$

- This all looks kosher and it works well! But...

  - Not a MLE; unclear what this optimises
  - Provably *inconsistent* and *biased*
  - Not actually ‘training’ the phrase-based model
Phrase-based Models

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Although north wind howls, but sky still very clear.
Phrase-based Models

Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。

However
Although north wind howls, but sky still very clear.

虽然北风呼啸，但天空依然十分清澈。
Although north wind howls, but sky still very clear.

However, the strong north wind, the sky remained clear under.

Thursday, 24 November 2011
Although north wind howls, but sky still very clear.

However, the strong north wind, the sky remained clear under.

Phrase-based Models
Although north wind howls, but sky still very clear.

Although north wind howls, but sky still very clear.

However, the sky remained clear under the strong north wind.

However, the sky remained clear under the strong north wind.
Although north wind howls, but sky still very clear.

\[
p(\text{English, alignment}|\text{Chinese}) = p(\text{segmentation}) \cdot p(\text{translations}) \cdot p(\text{reorderings})
\]

Thursday, 24 November 2011
Phrase-based Models
Phrase-based Models

- Segmentation probabilities.
Phrase-based Models

• Segmentation probabilities.
• Phrase translation probabilities.
Phrase-based Models

- Segmentation probabilities.
- Phrase translation probabilities.
- Distortion probabilities.
Phrase-based Models

- Segmentation probabilities.
- Phrase translation probabilities.
- Distortion probabilities.
Phrase-based Models

• Segmentation probabilities.
• Phrase translation probabilities.
• Distortion probabilities.

• Some problems:
  • Weak reordering model -- output is not fluent.
  • Many decisions -- many things can go wrong.
Carlos García has three associates.

García y asociados.

His associates are not strong.

Sus asociados no son fuertes.

García has a company also.

García también tiene una empresa.

Its clients are angry.

Sus clientes están enfadados.

The associates are also angry.

Los asociados también están enfadados.

The company has three groups.

La empresa tiene tres grupos.

Its groups are in Europe.

Sus grupos están en Europa.

The modern groups sell strong pharmaceuticals.

Los grupos modernos venden medicinas fuertes.

The groups do not sell zanzanina.

Los grupos no venden zanzanina.

The small groups are not modern.

Los grupos pequeños no son modernos.
Garcia and associates.

his associates are not strong.

Garcia has a company also.

its clients are angry.

the associates are also angry.

the company has strong enemies in Europe.

the clients and the associates are enemies.

los asociados tambien estan enfadados.

la empresa tiene enemigos fuertes en Europa.

los clientes y los asociados son enemigos.

the company has three groups.

its groups are in Europe.

los grupos tambien estan enfadados.

los asociados tambien estan enfadados.

los grupos pequenos no son modernos.

Warning: these are not phrases in any linguistic sense.
Warning: these are not phrases in any linguistic sense.

Restriction to linguistic phrases seems to hurt (Koehn et al. 2003)
Phrase-based Models
Phrase-based Models

- Phrase-based Models are dumb.
Phrase-based Models

• Phrase-based Models are dumb.
• But they are widely regarded as state-of-the-art.
Phrase-based Models

• Phrase-based Models are dumb.
• But they are widely regarded as state-of-the-art.
• Why? Simple models are easier to learn and deploy.

Thursday, 24 November 2011
Phrase-based Models

• Phrase-based Models are dumb.
• But they are widely regarded as state-of-the-art.
• Why? Simple models are easier to learn and deploy.
• Need proof? Google uses a phrase-based model.
However, the sky remained clear under the strong north wind.
Overview

However, the sky remained clear under the strong north wind.
Decoding

Probability models enable us to *make predictions*: Given a particular Chinese sentence, what is the most probable English sentence corresponding to it?
Decoding

Probability models enable us to *make predictions*: Given a particular Chinese sentence, what is the most probable English sentence corresponding to it?

In math, we want to solve:

$$\text{argmax}_{\text{English}} p(\text{English} | \text{Chinese})$$
Decoding

Probability models enable us to *make predictions*: Given a particular Chinese sentence, what is the most probable English sentence corresponding to it?

In math, we want to solve:

\[
\operatorname{argmax}_{\text{English}} p(\text{English}|\text{Chinese})
\]

problem: there are a lot of English sentences to choose from!
北风呼啸。
北风呼啸。

segmentations
substitutions
permutations
segmentations $O(2^n)$
substitutions
permutations
北 风 呼啸。

segmentations $O(2^n)$
substitutions $O(5^n)$
permutations
segmentations $O(2^n)$
substitutions $O(5^n)$
permutations $O(n!)$
segmentations $O(2^n)$
substitutions $O(5^n)$
permutations $O(n!)$

240,000 possibilities!
北风呼啸。
the strong north wind.
the strong north wind.
Given a sentence pair and an alignment, we can easily calculate

\[ p(\text{English, alignment}|\text{Chinese}) \]
Given a sentence pair and an alignment, we can easily calculate
\[ p(\text{English, alignment}|\text{Chinese}) \]

Can we do this without enumerating \( O(10^n n!) \) pairs?
Key Idea

北风呼啸。

Thursday, 24 November 2011
Key Idea

the strong north wind.
Key Idea

the strong north wind.
Key Idea

北风呼啸。

the strong north wind.

There are $O(10^n n!)$ target sentences.
Key Idea

There are $O(10^n n!)$ target sentences.

But there are only $O(5n^2)$ ways to start them.
Key Idea

北风呼啸。
Key Idea
Key Idea

coverage vector
Key Idea

coverage vector
Key Idea

\[ p(\text{north}|\text{START}) \cdot p(\text{北}|\text{north}) \]

coverage vector

北 风 呼啸。
Key Idea

$p(north|START) \cdot p(北|north)$

coverage vector

Thursday, 24 November 2011
Key Idea

\[ p(\text{north}|\text{START}) \cdot p(\text{北}|\text{north}) \]

\[ p(\text{northern}|\text{START}) \cdot p(\text{北}|\text{northern}) \]

coverage vector

Thursday, 24 November 2011
Key Idea

Coverage vector

\[ p(north | START) \cdot p(北 | north) \]

\[ p(northern | START) \cdot p(北 | northern) \]

强

北部
Key Idea

\[ p(\text{strong}|\text{START}) \cdot p(\text{北}|\text{strong}) \]

\[ p(\text{north}|\text{START}) \cdot p(\text{北}|\text{north}) \]

\[ p(\text{northern}|\text{START}) \cdot p(\text{北}|\text{northern}) \]

coverage vector
Key Idea

\[ p(north|\text{START}) \cdot p(\text{北}|north) \]

coverage vector
Key Idea

\[ p(\text{north}|\text{START}) \cdot p(\text{北}|\text{north}) \]

north → wind

coverage vector

北 风 呼啸。
Key Idea

\[ p(north|START) \cdot p(北|north) \cdot p(wind|north) \cdot p(风|wind) \]

coverage vector


Thursday, 24 November 2011
Key Idea

$p(north|START) \cdot p(\text{北}|north)$

$p(wind|north) \cdot p(\text{风}|wind)$

coverage vector

北 风 呼啸。
Key Idea

$p(north|START) \cdot p(北|north)$

$p(wind|north) \cdot p(风|wind)$

$p(strong|north) \cdot p(呼啸|strong)$

coverage vector

Thursday, 24 November 2011
Key Idea

Work done at sentence beginnings is shared across many possible output sentences!

\[ p(\text{north}|\text{START}) \cdot p(\text{北}|\text{north}) \]

\[ p(\text{wind}|\text{north}) \cdot p(\text{风}|\text{wind}) \]

\[ p(\text{strong}|\text{north}) \cdot p(\text{呼啸}|\text{strong}) \]
Key Idea
Key Idea
Key Idea

Dynamic Programming

Thursday, 24 November 2011
amount of work:

\[ O(5n^2 2^n) \]
Key Idea

amount of work:

\[ O(5n^22^n) \]

bad, but much better than

\[ O(10^n n!) \]

Dynamic Programming
Key Idea

amount of work:

$O(5n^2 2^n)$

bad, but much better than

$O(10^n n!)$
Key Idea

amount of work: $O(5n^22^n)$

bad, but much better than $O(10^n n!)$

each edge labelled with a weight and a word (or words)

Dynamic Programming

Thursday, 24 November 2011
Key Idea

amount of work:
\[ O(5n^2 2^n) \]

bad, but much better than
\[ O(10^n n!) \]

each edge labelled with a weight and a word (or words)

Dynamic Programming
Key Idea

amount of work:
\[ O(5n^2 2^n) \]

bad, but much better than
\[ O(10^n n!) \]

weighted finite-state automata

north, 0.014

each edge labelled with a weight and a word (or words)

Dynamic Programming

Thursday, 24 November 2011
Weighted languages

• The lattice describing the set of all possible translations is a *weighted finite state automaton*.

• So is the language model.

• Since regular languages are closed under intersection, we can intersect the devices and run shortest path graph algorithms.

• Taking their intersection is equivalent to computing the probability under Bayes’ rule.
虽然北风呼啸，但天空依然十分清澈。
虽然北风呼啸，但天空依然十分清澈。
START Although

虽然北风呼啸，但天空依然十分清澈。

crystal clear

虽然北风呼啸，但天空依然十分清澈。
Although crystal clear

Thursday, 24 November 2011
wind screamed

虽然北风呼啸，但天空依然十分清澈。

wind shrieked

虽然北风呼啸，但天空依然十分清澈。

north wind

虽然北风呼啸，但天空依然十分清澈。
wind screamed

虽然北风呼啸，但天空依然十分清澈。

wind shrieked

虽然北风呼啸，但天空依然十分清澈。

north wind

虽然北风呼啸，但天空依然十分清澈。
shrieked,

the sky

, yet

Thursday, 24 November 2011
the sky, yet

Although the wind shrieked, the sky remained clear.
虽然北风呼啸，但天空依然十分清澈。
虽然北风呼啸，但天空依然十分清澈。
still quite clear.

虽然北风呼啸, 但天空依然十分清澈。

blue.

虽然北风呼啸, 但天空依然十分清澈。
still quite clear.

blue.

Thursday, 24 November 2011
Although the northern wind shrieked across the sky, but was still very clear.

虽然北风呼啸，但天空依然十分清澈。
Practical Issues
Practical Issues

\[ O(5n^2 2^n) \] is still far too much work.
Practical Issues

$O(5n^22^n)$ is still far too much work.

Can we do better?
Practical Issues

$O(5n^22^n)$ is still far too much work.

Can we do better?

NO! Knight (1999) shows that this is NP-Complete.
Approximation: Pruning
Approximation: Pruning

Idea: prune states by accumulated path length
Approximation: Pruning
Approximation: Pruning

reality: longer paths have lower probability!
Approximation: Pruning

Solution: Group states by number of covered words.
Approximation: Pruning

Solution: Group states by number of covered words.
Approximation: Pruning

Solution: Group states by number of covered words.
Approximation: Pruning

“Stack” decoding: a linear-time approximation
Approximation: Distortion Limits

the sky

虽然北风呼啸，但天空依然十分清澈。
Approximation: Distortion Limits

number of vertices: $O(2^n)$
Approximation: Distortion Limits

number of vertices: \( O(2^n) \)

d = 4

window
Approximation: Distortion Limits

number of vertices: \( O(2^n) \)

outside window
to left: covered

\[ d = 4 \]

outside window
to right: uncovered
Approximation: Distortion Limits

number of vertices: $O(n2^d)$

d = 4

outside window to left: covered

outside window to right: uncovered

虽然北风呼啸，但天空依然十分清澈。
Interim Summary
Interim Summary

- We need every possible trick to make decoding fast.
Interim Summary

- We need every possible trick to make decoding fast.
- Dynamic programming: greatly reduces complexity of exact search, but still too slow.
Interim Summary

- We need every possible trick to make decoding fast.
- Dynamic programming: greatly reduces complexity of exact search, but still too slow.
- NP-Completeness means exact solutions unlikely.
Interim Summary

• We need every possible trick to make decoding fast.

• Dynamic programming: greatly reduces complexity of exact search, but still too slow.

• NP-Completeness means exact solutions unlikely.

• Common approximations: stack decoding, distortion limits
Interim Summary

- We need every possible trick to make decoding fast.
- Dynamic programming: greatly reduces complexity of exact search, but still too slow.
- NP-Completeness means exact solutions unlikely.
- Common approximations: stack decoding, distortion limits
- But, these approximations have a cost: we may not find the true argmax.
Implementations

- Phrase-based Translation
  - Moses -- [www.statmt.org/moses/](http://www.statmt.org/moses/)
  - cdec -- [www.cdec-decoder.org](http://www.cdec-decoder.org)

- Language models
  - RandLM -- [sourceforge.net/projects/randlm](http://sourceforge.net/projects/randlm)
Recap: Finite-State Models
Recap: Finite-State Models

• Probability theory enables us to learn from data.
Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!
Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!
- There’s no data like more data.
Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!
- There's no data like more data.
- Word-based models follow intuitions, but not all.
Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!
- There’s no data like more data.
- Word-based models follow intuitions, but not all.
- Phrase-based models are similar, but more effective.
Recap: Finite-State Models

• Probability theory enables us to learn from data.
• Very simple models get us pretty far!
• There’s no data like more data.
• Word-based models follow intuitions, but not all.
• Phrase-based models are similar, but more effective.
• *All* of these models are weighted regular languages.
Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!
- There’s no data like more data.
- Word-based models follow intuitions, but not all.
- Phrase-based models are similar, but more effective.
- *All* of these models are weighted regular languages.
- Need dynamic programming with approximations.
Recap: Finite-State Models

- Probability theory enables us to learn from data.
- Very simple models get us pretty far!
- There’s no data like more data.
- Word-based models follow intuitions, but not all.
- Phrase-based models are similar, but more effective.
- All of these models are weighted regular languages.
- Need dynamic programming with approximations.
- Is this the best we can do?
ارتفاع عجز الميزان التجاري الأردني

أفادت بيانات رسمية بأن العجز في الميزان التجاري الأردني ارتفع في النصف الأول من العام الحالي بنسبة 18.1% نتيجة زيادة حجم الواردات مقابل الصادرات، كما تراجع حجم القروض الائتمانية التي قدمتها البنوك الأردنية بنسبة 11% خلال الفترة نفسها.

وقالت بيانات دائرة الإحصاءات العامة نشرت اليوم الاثنين إن قيمة العجز في الميزان التجاري بلغت 2.79 مليار دينار أردني (3.94 مليار دولار أمريكي).

وأشارت البيانات إلى ارتفاع حجم الصادرات خلال النصف الأول من العام بنسبة 16.6%, حيث بلغت 2.05 مليار دينار (2.91 مليار دولار), كما ارتفعت قيمة الودادات بنسبة 11.7% لتبلغ 5.23 مليارات دينار (7.39 مليارات دولار).
Value of Jordanian imports amounted to 7.39 billion dollars in the first half of the year (island Net).

According to official statements that the trade balance deficit rose Jordan in the first half of this year by 18.1% due to increased volume of imports versus exports, as the decline in the volume of credit provided by banks of Jordan by 11% during the same period.

The data for the Department of Statistics published on Monday that the value of the trade balance deficit amounted to 2.79 billion Jordanian dinars (3.94 billion U.S. dollars). The data indicated the high volume of exports during the first half of the year by 16.6%, reaching 2.05 billion dinars (2.91 billion dollars), as imports rose by 11.7% to 5.23 billion dinars (7.39 billion dollars).

The export value increased significantly for the countries of the Greater Arab Free Trade, including Saudi Arabia, and the countries of the free trade agreement for North America, including the United States, as well as non-Arab Asian...
ปัจจุบันมีผู้ประกอบการที่ทำธุรกิจเกี่ยวกับปูยามาชิซึ่งทั่วถึงผู้ร้านซ่อม ผู้ผลิต และผู้จำหน่าย จึงจำเป็นต้องมีการขึ้นทะเบียนเพื่อตรวจสอบและรับรองคุณภาพจากกรมวิชาการเกษตร เพื่อแก้ปัญหา ปูยามาและปูยี ต่อคุณภาพ เป็นการปกป้องผลประโยชน์ให้กับพื้นที่เกษตรกร

นายสมชาย ชานฤทธิ์ กรมวิชาการเกษตรกล่าวว่า ปัจจุบันมีการดำเนินการธุรกิจเกี่ยวกับปูยีที่ไม่ถูกต้องสร้างความเสียหายให้กับเกษตรกร ทั้งนี้ปูยีที่จำหน่ายในท้องตลาดมี 2 ลักษณะ คือ 1. ปูยีบัลช์ แบบบอส (Bulk Blending) คือปูยีที่นำมาผสมกันเพื่อให้ได้สูตร ธาตุอาหารตามที่ใส่เข็นทะเบียนไว้แล้วน่า มากจาหน่าย 2. ปูยีเชิงเสีย คือ ปูยีที่เป็นเม็ดแล้วใต้มาตรฐานของธาตุอาหารตามที่ขึ้นทะเบียนไว้ เช่น ปูยีไพรเจน ปูยีฟาวร์ซ์ ปูยีโฟเทลซิซม
Organize ... enhance the quality fertilizer.
Tuesday, August 17, 2553 at 5:00 a.

There are currently entrepreneurs doing business with more fertilizer. One of the leading producer and marketer is required to verify the registration and certification from the Department of Agriculture To resolve fake fertilizers or fertilizers second Protect the benefits to farmers brothers.

Mr. Somchai Channarong Kun-General, Department of Agriculture said the current business operations on the wrong fertilizer damage to the farmers. The commercial fertilizer has two characteristics: 1. Fertilizer Bulk Blend (Bulk Blending) is the parent fertilizer fertilizer combinations to get the formula Nutrients as a registered already be sold 2. Polycultural fertilizer is fertilizer tablets and statues of nutrient standards for registration as the nitrogen fertilizer such as potassium, phosphorus fertilizer.
Learning the phrase-based model
Noisy Channel Model

\[ p(\text{English}|\text{Chinese}) \sim p(\text{English}) \times p(\text{Chinese}|\text{English}) \]

Two components, the language model and the translation model, both weighted equally.
Phrase-based model

- Has a language model and a translation model...
- But what about
  - the reordering model to penalise long jumps
  - the segmentation model over the chunking of the Chinese
- How should these be weighted?

\[
p(\text{English}|\text{Chinese}) \sim \]

\[
p(\text{English}) \times p(\text{Chinese}|\text{English}) \times p(\text{segmentation}) \times p(\text{reordering})
\]
Phrase-based model

• Has a *language model* and a *translation model*...

• But what about

  • the *reordering model* to penalise long jumps
  
  • the *segmentation model* over the chunking of the Chinese

• How should these be weighted?

\[
p(English|Chinese) \sim \lambda_1 p(English) \times \lambda_2 p(Chinese|English) \times \lambda_3 p(segmentation) \times \lambda_4 p(reordering)
\]
While we’re at it...

- Why not add some more features?
  - word count
  - phrase count
  - forward translation, $p(\text{English} \mid \text{Chinese})$
  - word-based translation, e.g., under model 1
- Now we just need to estimate the $\lambda$ values...
\[ p(English|Chinese) = \frac{1}{Z} \exp \left[ \sum_i \lambda_i h_i(Chinese, English) \right] \]
\[
\argmax_{English} \quad p(English | Chinese) = \frac{1}{Z} \exp \left[ \sum_i \lambda_i h_i(Chinese, English) \right]
\]
argmax \hspace{1em} p(English|Chinese) = \\
\frac{1}{Z} \exp \left[ \sum_{i} \lambda_i h_i(Chinese, English) \right]
\[
\arg\max_{English} \ p(English \mid Chinese) = \frac{1}{Z} \ \exp \left[ \sum_i \lambda_i h_i(Chinese, English) \right]
\]
\[
\text{argmax } p(\text{English}|\text{Chinese}) = \\
\text{English}
\]

\[
\left[ \sum_i \lambda_i h_i(\text{Chinese}, \text{English}) \right]
\]

linear model
Linear Models
Linear Models

• We could optimize lambdas for likelihood (this would be a log-linear model).
Linear Models

• We could optimize lambdas for likelihood (this would be a log-linear model).

• Good news: optimization is convex.
Linear Models

- We could optimize lambdas for likelihood (this would be a log-linear model).
- Good news: optimization is convex.
- Bad news: computing $Z$ is intractable (more on this later for syntax-based models).
Linear Models

- We could optimize lambdas for likelihood (this would be a log-linear model).

- Good news: optimization is convex.

- Bad news: computing $Z$ is intractable (more on this later for syntax-based models).

- Question: why should we bother with likelihood?
BLEU\((MT \ text)\)
BLEU(\(\text{argmax} \ score(English|Chinese)\))

English
\[
\sum_{\text{Chinese} \in \text{Test}} \text{BLEU} \left( \text{argmax } \text{score} \left( \frac{\text{English}}{\text{Chinese}} \right) \right)
\]
Optimising for BLEU

Notice:

\[ \text{score}(\text{English} | \text{Chinese}) = \sum_{i} \lambda_i h_i(\text{Chinese, English}) \]
Optimising for BLEU

Notice:

\[ \text{score}(\text{English}|\text{Chinese}) = \lambda_x h_x(\text{Chinese, English}) + \sum_{i/x} \lambda_i h_i(\text{Chinese, English}) \]
Optimising for BLEU

Notice:

\[ score(English|Chinese) = a\lambda_x + b \]
Optimising for BLEU

Notice:

\[ \text{score}(\text{English}|\text{Chinese}) = a\lambda_x + b \]

just a line!
Optimising for BLEU

\[
\text{model score} \quad \lambda_x
\]
Optimising for BLEU

\[ \lambda \]

model score

\[ \lambda_x \]
Optimising for BLEU

![Graph showing the relationship between model score and $\lambda_x$.]
Optimising for BLEU

model score

$\lambda_x$

Thursday, 24 November 2011
Optimising for BLEU
Optimising for BLEU

model score

$\lambda_x$
Optimising for BLEU

model score

\( \lambda_x \)
Optimising for BLEU
Optimising for BLEU

model score

$\lambda_x$
Optimising for BLEU
Optimising for BLEU

BLEU (n-axis) vs $\lambda_x$ (x-axis)

Model score (y-axis) vs $\lambda_x$ (x-axis)
Optimising for BLEU
Optimising for BLEU

Minimum Error Rate Training (Och 2003)
Optimising for BLEU
Optimising for BLEU
Optimising for BLEU
Optimising for BLEU
Optimising for BLEU
Optimising for BLEU
Optimising for BLEU
Optimising for BLEU
Optimising for BLEU
Build a State-of-the-Art System

- Phrase-based models (or hierarchical or syntactic)
- on masses of parallel data
- Really really big language model
- Minimum error rate training
Overview

training data
(parallel text)

learner

model

decoder

However, the sky remained clear under the strong north wind.
Two Problems

• Exact decoding requires exponential time.
  • This is a consequence of arbitrary permutation.
  • But in translation reordering is not arbitrary!

• Parameterization of reordering is weak.
  • No generalization!
Garcia and associates.

Carlos García tiene tres asociados.

His associates are not strong.

Garcia has a company also.

Garcia también tiene una empresa.

Its clients are angry.

The associates are also angry.

The company has strong enemies in Europe.

The clients and the associates are enemies.

The company has three groups.

Its groups are in Europe.

The modern groups sell strong pharmaceuticals.

The groups do not sell zanzanine.

The small groups are not modern.
la empresa tiene enemigos fuertes en Europa.

the company has **strong enemies** in Europe.

Garcia and associates.

Carlos García has three associates.

his associates are not strong.

Garcia has a company also.

the clients and the associates are enemies.

its clients are angry.

the associates are also angry.

the company has three groups.

its groups are in Europe.

**modern groups** sell **strong pharmaceuticals**.

the groups do not sell zanzanine.

**small groups** are not modern.

la empresa tiene tres grupos.

sus asociados no son fuertes.

Garcia también tiene una empresa.

sus clientes estan enfadados.

los asociados tambien estan enfadados.

los grupos modernos venden medicinas fuertes.

los grupos no venden zanzanina.

los grupos pequenos no son modernos.
Garcia and associates.

Carlos Garcia has three associates.

His associates are not strong.

Garcia has a company also.

The company has strong enemies in Europe.

The clients and the associates are enemies.

The modern groups sell strong pharmaceuticals.

The groups do not sell zanzanine.

The small groups are not modern.

Same pattern:

NN JJ \rightarrow JJ NN
la empresa tiene enemigos fuertes en Europa.

the company has strong enemies in Europe.

Same pattern:

NN JJ → JJ NN

Finite-state models do not capture this generalization.
Summary

• Many probabilistic translation models can be thought in terms of weighted (formal) languages.

• Dynamic programming is a common (though not universal!) decoding strategy.

• With these concepts in mind, you might be able to define models that capture other translation phenomena (e.g. morphosyntactics, semantics).