UNIVERSITY OF CAMBRIDGE

Bayesian Nonparametrics in Real-World Applications: Statistical Machine Translation and Language Modelling on Big Datasets

Yarin Gal

yg279@cam.ac.uk

16th of May 2013

Based in part on the lecture notes by Dr. Phil Blunsom

Outline

Outline

Introduction

Parallel corpora

Models of translation

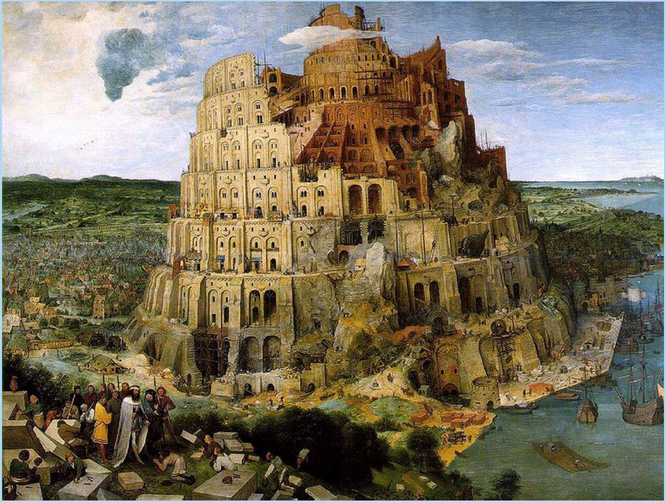
Word Alignment

Basic Bayesian approaches

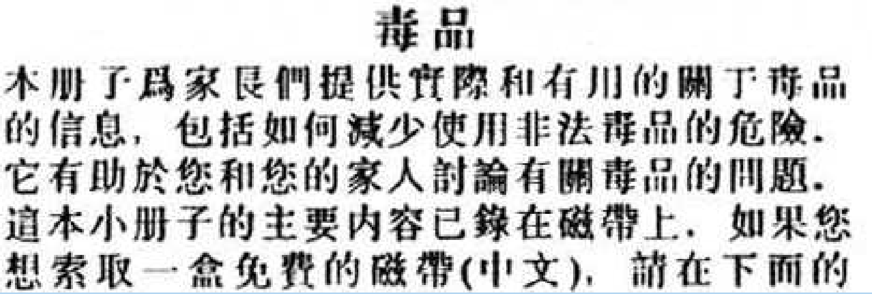
Bayesian Nonparametric approaches

Conclusions

The confusion of tongues:



Task: make sense of foreign text like:



毒品

本冊子爲家長們提供實際和有用的關于毒品
的信息，包括如何減少使用非法毒品的危險。
它有助於您和您的家人討論有關毒品的問題。
這本小冊子的主要內容已錄在磁帶上。如果您
想索取一盒免費的磁帶(中文)，請在下面的

- AI-hard: ultimately reasoning and world knowledge required
- Statistical machine translation: Learn how to translate from data

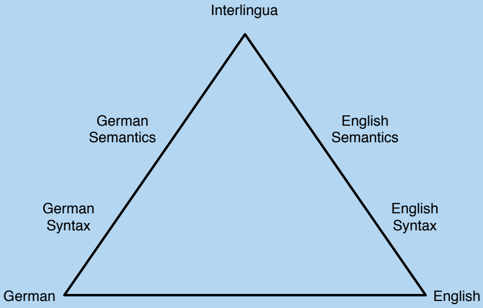
UNIVERSITY OF CAMBRIDGE

Warren Weaver memorandum, July 1949:

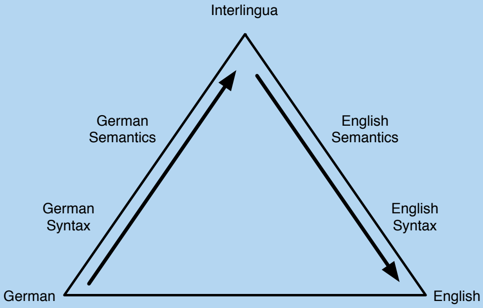
Thus it may be true that the way to translate from Chinese to Arabic, or from Russian to Portuguese, is not to attempt the direct route, shouting from tower to tower. Perhaps the way is to descend, from each language, down to the common base of human communication—the real but as yet undiscovered universal language—and—then re-emerge by whatever particular route is convenient.

UNIVERSITY OF CAMBRIDGE

The Machine Translation Pyramid:

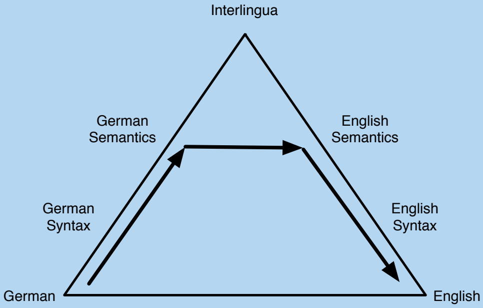


The Machine Translation Pyramid:

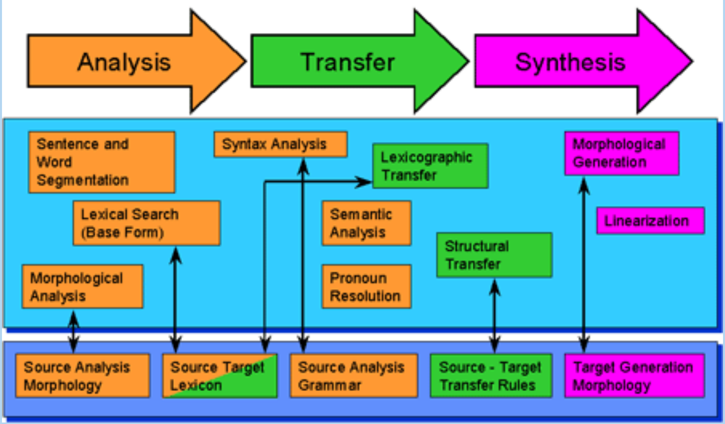


UNIVERSITY OF CAMBRIDGE

The Machine Translation Pyramid:



Rule Based Machine Translation (RBMT):



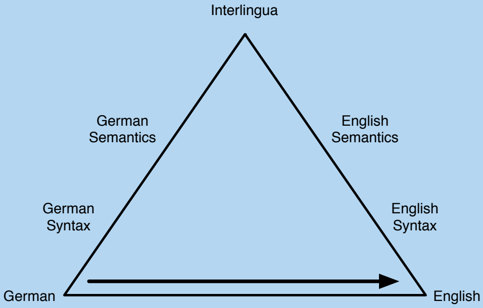
taken from `www.linguatec.net`

UNIVERSITY OF CAMBRIDGE

Warren Weaver memorandum, July 1949:

It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code." If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?

The Machine Translation Pyramid:



Fred Jelinek, 1988:

Every time I fire a linguist, the performance of the recognizer goes up.

Outline

Introduction

Parallel corpora

Models of translation

Word Alignment

Basic Bayesian approaches

Bayesian Nonparametric approaches

Conclusions

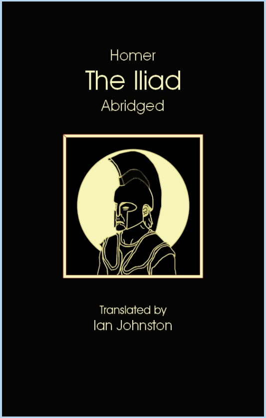
UNIVERSITY OF CAMBRIDGE

Rosetta Stone:

Parallel Corpora

UNIVERSITY OF CAMBRIDGE

Iliad:



UNIVERSITY OF CAMBRIDGE

UN Website:

Outline

UNIVERSITY OF CAMBRIDGE

Given an input sentence, we have to predict an output translation

Natuerlich hat John spass am Spiel.

$\Downarrow$

Of course John has fun with the game.

- Since the set of possible output sentences is too large, we need to construct the translation according to some decomposition of the translation process

UNIVERSITY OF CAMBRIDGE

The Noisy Channel Model

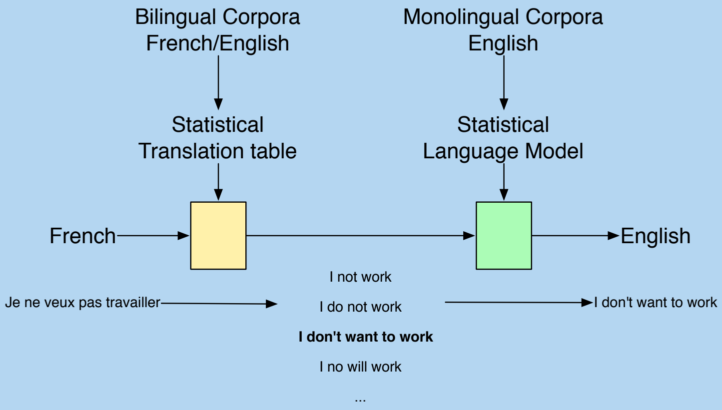
$$P(English|French) = \frac{P(English) \times P(French|English)}{P(French)}$$

$$\arg\max_{\mathbf{e}} P(\mathbf{e}|\mathbf{f}) = \arg\max_{\mathbf{e}} [P(\mathbf{e}) \times P(\mathbf{f}|\mathbf{e})]$$

- Bayes' rule is used to reverse the translation probabilities
- the analogy is that the French is English transmitted over a *noisy channel*
- we can then use techniques from statistical signal processing and decryption to translate

UNIVERSITY OF CAMBRIDGE

The Noisy Channel Model



UNIVERSITY OF CAMBRIDGE

(Bi-gram) Language Modelling

$$P(\mathbf{e}) = P(e_0, e_1, \ldots, e_{|\mathbf{e}|})$$
$$\approx \prod_{i=0}^{|\mathbf{e}|} P(e_i|e_{i-1})$$

- We can approximate the probability of seeing an English word e_i conditioned only on the previous word e_{i-1}.
- These conditional probabilities can be estimated from monolingual corpora.

UNIVERSITY OF CAMBRIDGE

(Bi-gram) Language Modelling: The Iliad

Sing, O goddess, the anger of Achilles son of Peleus, that brought countless ills upon the Achaeans. Many a brave soul did it send hurrying down to Hades, and many a hero did it yield a prey to dogs and vultures, for so were the counsels of Jove fulfilled from the day on which the son of Atreus, king of men, and great Achilles, first fell out with one another.

- $P(x|y) = \frac{count(x,y)}{count(y)}$

UNIVERSITY OF CAMBRIDGE

(Bi-gram) Language Modelling: The Iliad

Sing, O goddess, the anger of Achilles son of Peleus, that brought countless ills upon the Achaeans. Many a brave soul did it send hurrying down to Hades, and many a hero did it yield a prey to dogs and vultures, for so were the counsels of Jove fulfilled from the day on which the son of Atreus, king of men, and great Achilles, first fell out with one another.

- $P(x|the) = \frac{count(x, the)}{count(the)}$
- $P(son|the) = \frac{count(son, the)}{count(the)} = \frac{1}{5}$

UNIVERSITY OF CAMBRIDGE

(Bi-gram) Language Modelling: The Iliad

Sing, O goddess, the anger of Achilles son of Peleus, that brought countless ills upon the Achaeans. Many a brave soul did it send hurrying down to Hades, and many a hero did it yield a prey to dogs and vultures, for so were the counsels of Jove fulfilled from the day on which the son of Atreus, king of men, and great Achilles, first fell out with one another.

- $P(x|the) = \frac{count(x, the)}{count(the)}$
- $P(son|the) = \frac{count(son, the)}{count(the)} = \frac{1}{5}$
- $P(king|the) = \frac{count(king, the)}{count(the)} = \frac{0}{5}$?

UNIVERSITY OF CAMBRIDGE

Solution: smoothing and interpolation with shorter sequences of words

Interpolated Kneser-Ney discounting language model

$\mathbf{u}$ - a sequence of l words
$c_{\mathbf{u}w}$ - number of observations of the sequence $\mathbf{u}$ followed by the word w
$\pi(\mathbf{u})$ - $\mathbf{u}$ without the left most word

$$P_{\mathbf{u}}(w) = \frac{\max(0, c_{\mathbf{u}w} - d_l)}{c_{\mathbf{u}.}} + \frac{d_l t_{\mathbf{u}.}}{c_{\mathbf{u}.}} P_{\pi(\mathbf{u})}(w)$$

where $c_{\mathbf{u}.}$ is the number of observations of the sequence $\mathbf{u}$, $t_{\mathbf{u}.}$ is the number of unique words following the sequence $\mathbf{u}$, $P_{\emptyset}$ is a uniform distribution over all words, and d_l depends on length l.

- Shorter sequences of words will have higher weight in the interpolation if $\mathbf{u}w$ is sparse

UNIVERSITY OF CAMBRIDGE

(Bi-gram) Interpolated Kneser-Ney: The Iliad

Sing, O goddess, the anger of Achilles son of Peleus, that brought countless ills upon the Achaeans. Many a brave soul did it send hurrying down to Hades, and many a hero did it yield a prey to dogs and vultures, for so were the counsels of Jove fulfilled from the day on which the son of Atreus, king of men, and great Achilles, first fell out with one another.

- $P_y(x) = \frac{\max(0, c_{y,x} - d_1)}{c_{y\cdot}} + \frac{d_1 t_{y\cdot}}{c_{y\cdot}} P_{\pi(y)}(x)$

UNIVERSITY OF CAMBRIDGE

(Bi-gram) Interpolated Kneser-Ney: The Iliad

Sing, O goddess, the anger of Achilles son of Peleus, that brought countless ills upon the Achaeans. Many a brave soul did it send hurrying down to Hades, and many a hero did it yield a prey to dogs and vultures, for so were the counsels of Jove fulfilled from the day on which the son of Atreus, king of men, and great Achilles, first fell out with one another.

- $P_{the}(x) = \frac{\max(0, c_{the,x} - d_1)}{c_{the\bullet}} + \frac{d_1 t_{the\bullet}}{c_{the\bullet}} P_{\pi(the)}(x)$
- $P_{the}(son) = \frac{\max(0, c_{the,son} - d_1)}{c_{the\bullet}} + \frac{d_1 t_{the\bullet}}{c_{the\bullet}} P_{\pi(the)}(son)$

UNIVERSITY OF CAMBRIDGE

(Bi-gram) Interpolated Kneser-Ney: The Iliad

Sing, O goddess, the anger of Achilles son of Peleus, that brought countless ills upon the Achaeans. Many a brave soul did it send hurrying down to Hades, and many a hero did it yield a prey to dogs and vultures, for so were the counsels of Jove fulfilled from the day on which the son of Atreus, king of men, and great Achilles, first fell out with one another.

- $P_{the}(x) = \frac{\max(0, c_{the,x} - d_1)}{c_{the\bullet}} + \frac{d_1 \, t_{the\bullet}}{c_{the\bullet}} P_{\pi(the)}(x)$
- $P_{the}(son) = \frac{1-d_1}{5} + \frac{5d_1}{5} P_{\epsilon}(son)$

UNIVERSITY OF CAMBRIDGE

(Bi-gram) Interpolated Kneser-Ney: The Iliad

Sing, O goddess, the anger of Achilles son of Peleus, that brought countless ills upon the Achaeans. Many a brave soul did it send hurrying down to Hades, and many a hero did it yield a prey to dogs and vultures, for so were the counsels of Jove fulfilled from the day on which the son of Atreus, king of men, and great Achilles, first fell out with one another.

- $P_{the}(x) = \frac{\max(0, c_{the,x} - d_1)}{c_{the\cdot}} + \frac{d_1 \, t_{the\cdot}}{c_{the\cdot}} P_{\pi(the)}(x)$
- $P_{the}(son) = \frac{1 - d_1}{5} + \frac{5 d_1}{5} \left(\frac{\max(0, c_{son} - d_0)}{c_{\cdot}} + P_{\emptyset}(son) \right)$

UNIVERSITY OF CAMBRIDGE

(Bi-gram) Interpolated Kneser-Ney: The Iliad

Sing, O goddess, the anger of Achilles son of Peleus, that brought countless ills upon the Achaeans. Many a brave soul did it send hurrying down to Hades, and many a hero did it yield a prey to dogs and vultures, for so were the counsels of Jove fulfilled from the day on which the son of Atreus, king of men, and great Achilles, first fell out with one another.

- $P_{the}(x) = \frac{\max(0, c_{the,x} - d_1)}{c_{the\cdot}} + \frac{d_1 \, t_{the\cdot}}{c_{the\cdot}} P_{\pi(the)}(x)$
- $P_{the}(son) = \frac{1-d_1}{5} + \frac{5d_1}{5}\left(\frac{2-d_0}{72} + \frac{1}{T}\right)$

UNIVERSITY OF CAMBRIDGE

(Bi-gram) Interpolated Kneser-Ney: The Iliad

Sing, O goddess, the anger of Achilles son of Peleus, that brought countless ills upon the Achaeans. Many a brave soul did it send hurrying down to Hades, and many a hero did it yield a prey to dogs and vultures, for so were the counsels of Jove fulfilled from the day on which the son of Atreus, king of men, and great Achilles, first fell out with one another.

- $P_{the}(x) = \frac{\max(0, c_{the,x} - d_1)}{c_{the\cdot}} + \frac{d_1 t_{the\cdot}}{c_{the\cdot}} P_{\pi(the)}(x)$
- $P_{the}(king) = \frac{0}{5} + \frac{5d_1}{5} P_{\epsilon}(king) = \frac{5d_1}{5}\left(\frac{1-d_0}{72} + \frac{1}{T}\right)$

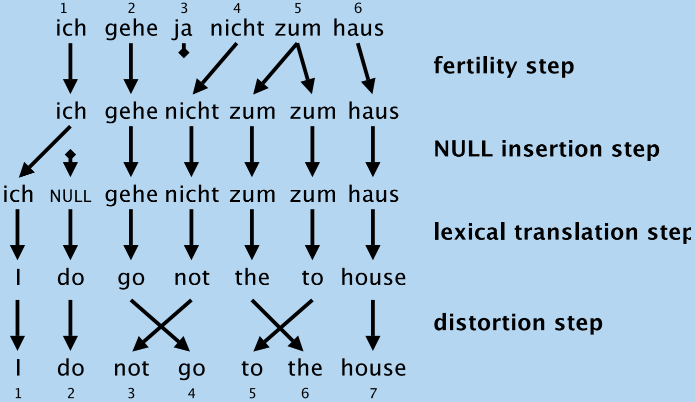
Language models

Language modelling in Machine Translation:

- very important!
- 5-gram models are now commonplace
- Such models require *lots* of data to estimate; we routinely use *billions* of words of English
- Smoothing is crucial for these higher order n-gram models

UNIVERSITY OF CAMBRIDGE

Word-based translation



Original statistical machine translation models (1990s): break down translation to the word level

UNIVERSITY OF CAMBRIDGE

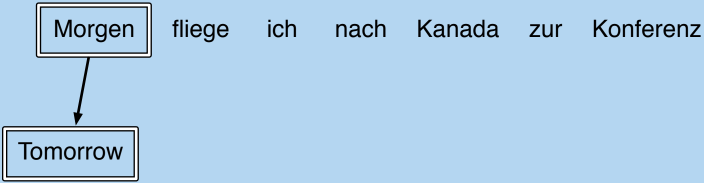
Phrase-based translation

Morgen fliege ich nach Kanada zur Konferenz

Current state of the art:
map larger chunks of words (huge mapping tables)

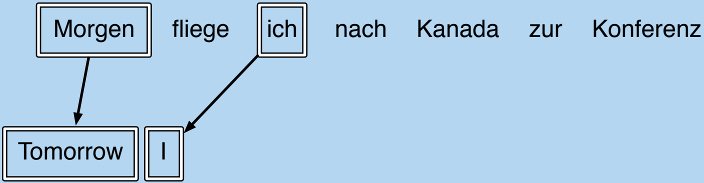
UNIVERSITY OF CAMBRIDGE

Phrase-based translation



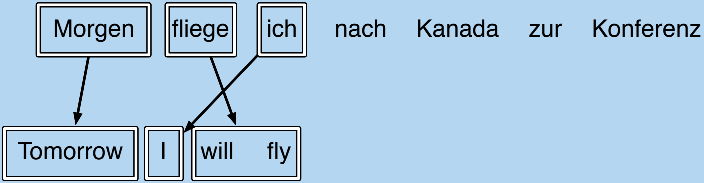
Current state of the art:
map larger chunks of words (huge mapping tables)

Phrase-based translation



Current state of the art:
map larger chunks of words (huge mapping tables)

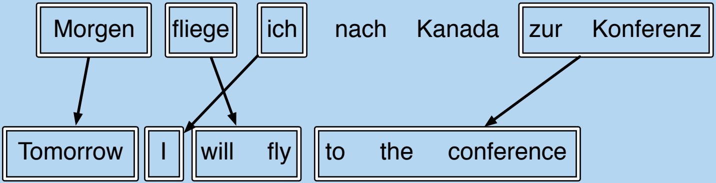
Phrase-based translation



Current state of the art:
map larger chunks of words (huge mapping tables)

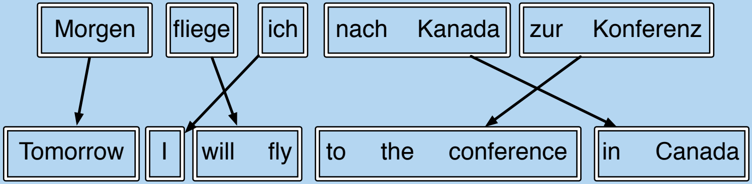
UNIVERSITY OF CAMBRIDGE

Phrase-based translation



Current state of the art:
map larger chunks of words (huge mapping tables)

Phrase-based translation



Current state of the art:
map larger chunks of words (huge mapping tables)

Advantages of phrase-based approach:

- improved modelling of multi-word translation units
- increased context
- permits idioms and non-compositional phrases
- eases search and reliance on the language model

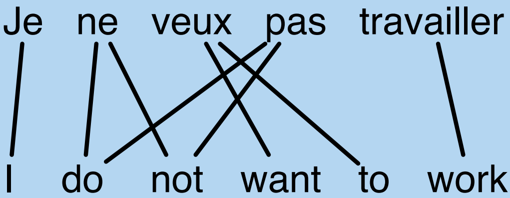
Phrase extraction:

Je ne veux pas travailler

I do not want to work

UNIVERSITY OF CAMBRIDGE

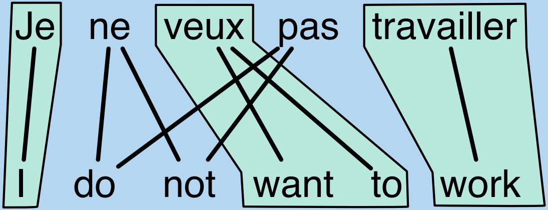
Phrase extraction:



- Use a word-based translation model to annotate the parallel corpus with word-alignments

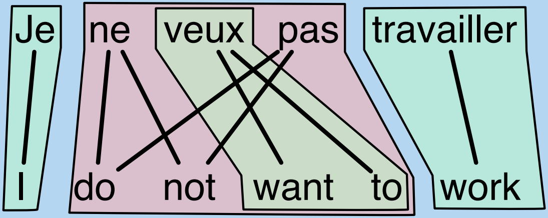
UNIVERSITY OF CAMBRIDGE

Phrase extraction:



- ⟨ Je, I ⟩, ⟨ veux, want to ⟩, ⟨ travailler, work ⟩

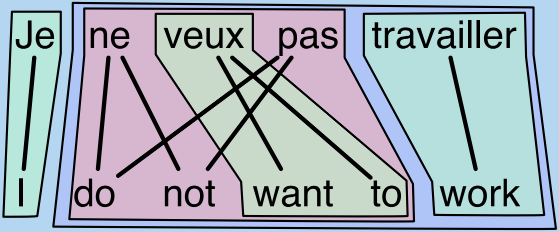
Phrase extraction:



- ⟨ Je, I ⟩, ⟨ veux, want to ⟩, ⟨ travailler, work ⟩, ⟨ ne veux pas, do not want to ⟩

UNIVERSITY OF CAMBRIDGE

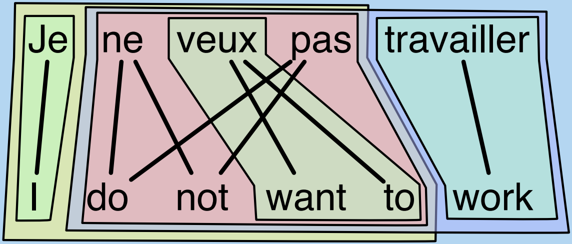
Phrase extraction:



- ⟨ Je, I ⟩, ⟨ veux, want to ⟩, ⟨ travailler, work ⟩, ⟨ ne veux pas, do not want to ⟩, ⟨ ne veux pas travailler, do not want to work ⟩

UNIVERSITY OF CAMBRIDGE

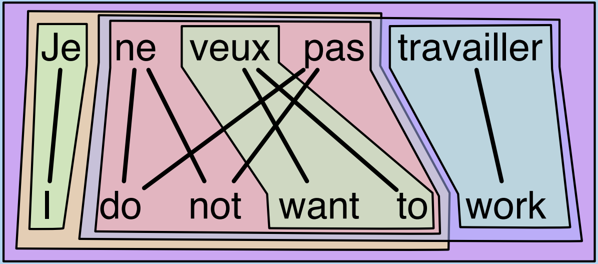
Phrase extraction:



- ⟨ Je, I ⟩, ⟨ veux, want to ⟩, ⟨ travailler, work ⟩, ⟨ ne veux pas, do not want to ⟩, ⟨ ne veux pas travailler, do not want to work ⟩, ⟨ Je ne veux pas, I do not want to ⟩

UNIVERSITY OF CAMBRIDGE

Phrase extraction:



- ⟨ Je, I ⟩, ⟨ veux, want to ⟩, ⟨ travailler, work ⟩, ⟨ ne veux pas, do not want to ⟩, ⟨ ne veux pas travailler, do not want to work ⟩, ⟨ Je ne veux pas, I do not want to ⟩, ⟨ Je ne veux pas travailler, I do

Outline

UNIVERSITY OF CAMBRIDGE

Word Alignment (IBM Model 1)

UNIVERSITY OF CAMBRIDGE

A simple generative model for $p(F|E)$ is derived by introducing a latent variable A into the conditional probability:

$$p(F, A|E) = \frac{p(J|I)}{(I+1)^J} \prod_{j=1}^{J} p(f_j|e_{a_j}),$$

- F and E are the input (source) and output (target) sentences of length J and I respectively,
- A is a vector of length J consisting of integer indexes into the target sentence, known as the alignment.

To learn this model the EM algorithm is used to find the MLE values for the parameters $p(f_j|e_{a_j})$.

- For the EM update we need to calculate the conditional probability of an alignment $p(A|E, F) = \frac{p(F,A|E)}{p(F|E)}$

Word Alignment (IBM Model 1)

Marginalising out A in $p(F, A|E)$ gives the required denominator:

$$p(F|E) = \sum_{A} p(F, A|E)$$

$$= \sum_{a_1=0}^{I} \sum_{a_2=0}^{I} \cdots \sum_{a_J=0}^{I} p(F, A|E)$$

$$= \frac{p(J|I)}{(I+1)^J} \sum_{a_1=0}^{I} \sum_{a_2=0}^{I} \cdots \sum_{a_J=0}^{I} \prod_{j=1}^{J} p(f_j|e_{a_j})$$

Rather conveniently we can swap the sum and product in the last line to get an equation that is tractable to compute:

$$= \frac{p(J|I)}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(f_j|a_i).$$

The result is that we can calculate the counts in $\mathcal{O}(J \times I)$ rather than $\mathcal{O}(I+1)^J$.

Word Alignment (IBM Model 1)

Limitations of this simple word alignment model:

- The structure of sentences is not modelled, words align independently of each other,
- The position of words with a sentence is not modelled, obviously words near the start of the source sentence are more likely to align to words near the start of the target sentence,
- The alignment is asymmetric, a target word may align to multiple source words, but a source word may only align to a single target,
- and many others ...

These limitations mean that this model does not work well as a translation model on it's own, however it is currently used as the first step in learning more complicated models by online translation providers such as Google and Microsoft.

A more accurate generative model for $p(F|E)$ is derived by introducing dependencies between alignment positions.

The HMM alignment model

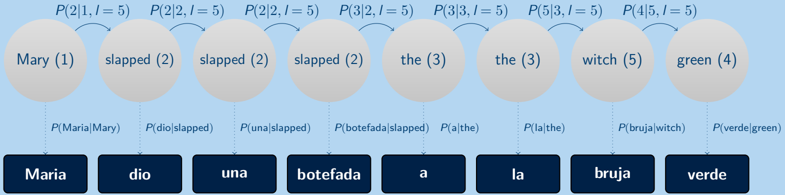
$$P(F, A|E) \quad = \quad P(l|m) \times \prod_{j=1}^{l} \left(P(a_j|a_{j-1}, l) \times P(f_j|e_{a_j}) \right)$$

- We can use the Forward-Backward algorithm for tractable training of this model

UNIVERSITY OF CAMBRIDGE

Alignment in this model can be found by jumping over the English sentence words and emitting foreign words.

HMM alignment model



Aligning the sentence pair (*"Mary slapped the green witch"*, *"Maria dio una bofetada a la bruja verde"*)

Word Alignment (Fertility based models)

Another class of alignment models is the fertility based models. These models follow more of a linguistic approach than the previous ones that used mathematical conveniences.

- We treat the alignment as a function from the source sentence positions i to $B_i \subset \{1, ..., m\}$ where the B_i's form a partition of the set $\{1, ..., m\}$,
- We define the fertility of the English word i to be $\phi_i = |B_i|$, the number of foreign words it generated,
- And $B_{i,k}$ refers to the kth word of B_i from left to right.

We allow for additional, spurious, words to be generated by introducing the NULL word at the beginning of the English sentence.

Probability model:

$$P(F, A|E) = p(B_0|B_1, ..., B_I) \times \prod_{i=1}^{I} p(B_i|B_{i-1}, e_i) \times \prod_{i=0}^{I} \prod_{j \in B_i} p(f_j|e_i)$$

2 main models belong to this class: IBM model 3 and IBM model 4. For model 3 the dependence on previous alignment sets is ignored and the probability $p(B_i|B_{i-1}, e_i)$ is modelled as

$$p(B_i|B_{i-1}, e_i) = p(\phi_i|e_i)\phi_i! \prod_{j \in B_i} p(j|i, m),$$

Whereas for model 4 it is modelled using two HMMs:

$$p(B_i|B_{i-1}, e_i) = p(\phi_i|e_i) \times p_{=1}(B_{i,1} - \odot(B_{i-1})|\cdot) \times \prod_{k=2}^{\phi_i} p_{>1}(B_{i,k} - B_{i,k-1}|\cdot)$$

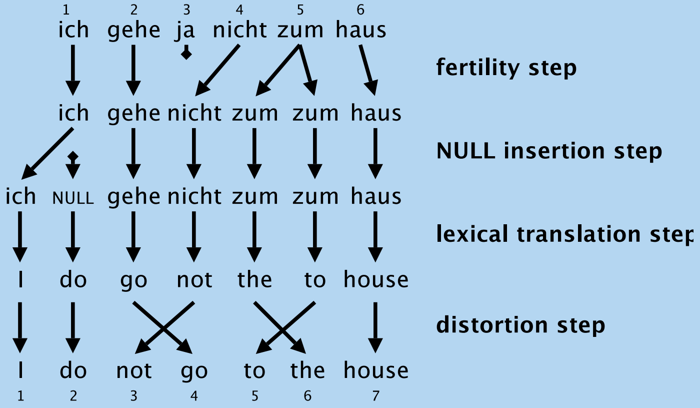
For both these models the spurious word generation is controlled by a binomial distribution:

$$p(B_0|B_1, ..., B_l) = \binom{m - \phi_0}{\phi_0} (1 - p_0)^{m - 2\phi_0} p_1^{\phi_0} \frac{1}{\phi_0!}$$

for some parameters p_0 and p_1.

UNIVERSITY OF CAMBRIDGE

Models 3 and 4 word alignment



Limitations:

- Inference in models 3 and 4 is intractable
  - We know of no efficient way to avoid the explicit summation over all alignments in the EM algorithm in the fertility-based alignment models
  - To circumvent this, the counts are collected only over a small neighbourhood of good alignments
  - To keep the training fast, we consider only a small fraction of all alignments.
- Sparsity is not handled

Outline

Basic Bayesian approaches

The alignment models mentioned have underpinned the majority of statistical machine translation systems for almost twenty years.

- They offer principled probabilistic formulation and (mostly) tractable inference
- There are many open source packages implementing them
  - Giza++ – one of the dominant implementations,
  - employs a variety of exact and approximate EM algorithms

However –

Basic Bayesian approaches

However –

- The parametric approach results in a significant number of parameters to be tuned
- Intractable summations over alignments for models 3 and 4
  - Usually approximated using restricted alignment neighbourhoods
  - Shown to return alignments with probabilities well below the true maxima
- Sparse contexts are not handled

Many alternative approaches to word alignment have been proposed, and largely failed to dislodge the IBM approach.

However –

- The parametric approach results in a significant number of parameters to be tuned
- Intractable summations over alignments for models 3 and 4
  - Usually approximated using restricted alignment neighbourhoods
  - Shown to return alignments with probabilities well below the true maxima
- Sparse contexts are not handled

Many alternative approaches to word alignment have been proposed, and largely failed to dislodge the IBM approach.

What can we do instead?

One possible solution:

Dirichlet prior

Put a Dirichlet prior over the categorical distribution for the word translation and alignment transition probabilities of the HMMs used in the different models:

$$\mathbf{t}_e \sim \mathit{Dir}(\Theta_e)$$

$$f_j|\mathbf{a}, \mathbf{e}, \mathbf{T} \sim \mathit{Categorical}(\mathbf{t}_{e_{a_j}})$$

$$\mathbf{a_j}|\alpha \sim \mathit{Dir}(\alpha)$$

$$a_{j+1}|a_j, \mathbf{a_j} \sim \mathit{Categorical}(\mathbf{a_j})$$

- Captures sparsity by using small values for the hyper-parameter

Basic Bayesian approaches

Several Bayesian inference mechanisms have also been recently adapted for the word alignment models training:

- Variational Bayes (2012)
- Collapse Variational Bayes (2013)

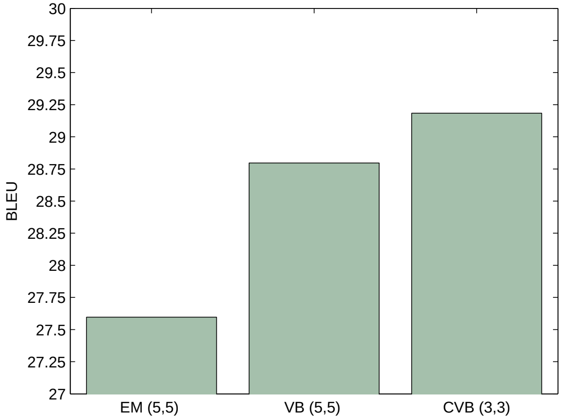
Using the BLEU metric, we can see the improvement in translation quality as more advanced techniques are used.

Several Bayesian inference mechanisms have also been recently adapted for the word alignment models training:

- Variational Bayes (2012)
- Collapse Variational Bayes (2013)

Using the BLEU metric, we can see the improvement in translation quality as more advanced techniques are used.

We can use the real-world application of Statistical Machine Translation for the assessment of different inference techniques!



BLEU scores of different systems translating from Chinese to English

Basic Bayesian approaches

Limitations of the variational approaches:

- Still have a significant number of parameters to tune
- In the case of models 3 and 4, still approximating using alignment neighbourhoods

Can we marginalise over all parameters?

Basic Bayesian approaches

Limitations of the variational approaches:

- Still have a significant number of parameters to tune
- In the case of models 3 and 4, still approximating using alignment neighbourhoods

Can we marginalise over all parameters?

- Gibbs sampling has been implemented in 2011 for IBM model 1 to use a fully Bayesian approach.

Basic Bayesian approaches

Limitations of the variational approaches:

- Still have a significant number of parameters to tune
- In the case of models 3 and 4, still approximating using alignment neighbourhoods

Can we marginalise over all parameters?

- Gibbs sampling has been implemented in 2011 for IBM model 1 to use a fully Bayesian approach.

We can still do better.

Outline

Bayesian Nonparametric approaches

Several smoothing techniques have been proposed in language modelling over the years.

- Add-one smoothing (1920) – adds one to all counts

Several smoothing techniques have been proposed in language modelling over the years.

- Add-one smoothing (1920) – adds one to all counts
- Good-Turing smoothing (1953) – improves over this by using the frequency of singletons to estimate the frequency of zero-count bigrams

Bayesian Nonparametric approaches

Several smoothing techniques have been proposed in language modelling over the years.

- Add-one smoothing (1920) – adds one to all counts
- Good-Turing smoothing (1953) – improves over this by using the frequency of singletons to estimate the frequency of zero-count bigrams
- Interpolated Kneser-Ney (1995) – a further improvement that includes absolute discounting

$$P_{\mathbf{u}}(w) = \frac{\max(0, c_{\mathbf{u}w} - d_{|\mathbf{u}|})}{c_{\mathbf{u}\cdot}} + \frac{d_{|\mathbf{u}|} t_{\mathbf{u}\cdot}}{c_{\mathbf{u}\cdot}} P_{\pi(\mathbf{u})}(w)$$

Bayesian Nonparametric approaches

Several smoothing techniques have been proposed in language modelling over the years.

- Add-one smoothing (1920) – adds one to all counts
- Good-Turing smoothing (1953) – improves over this by using the frequency of singletons to estimate the frequency of zero-count bigrams
- Interpolated Kneser-Ney (1995) – a further improvement that includes absolute discounting

$$P_{\mathbf{u}}(w) = \frac{\max(0, c_{\mathbf{u}w} - d_{|\mathbf{u}|})}{c_{\mathbf{u}\cdot}} + \frac{d_{|\mathbf{u}|} t_{\mathbf{u}\cdot}}{c_{\mathbf{u}\cdot}} P_{\pi(\mathbf{u})}(w)$$

  - one of the most commonly used modern N-gram smoothing methods in the NLP community

Bayesian Nonparametric approaches

Several smoothing techniques have been proposed in language modelling over the years.

- Add-one smoothing (1920) – adds one to all counts
- Good-Turing smoothing (1953) – improves over this by using the frequency of singletons to estimate the frequency of zero-count bigrams
- Interpolated Kneser-Ney (1995) – a further improvement that includes absolute discounting

$$P_{\mathbf{u}}(w) = \frac{\max(0, c_{\mathbf{u}w} - d_{|\mathbf{u}|})}{c_{\mathbf{u}\cdot}} + \frac{d_{|\mathbf{u}|} t_{\mathbf{u}\cdot}}{c_{\mathbf{u}\cdot}} P_{\pi(\mathbf{u})}(w)$$

  - one of the most commonly used modern N-gram smoothing methods in the NLP community
  - Was discovered in 2006 to corresponds exactly to a well-know stochastic process in the Bayesian Nonparametric community: the hierarchical Pitman-Yor process.

We can define the Pitman-Yor process by describing how to draw from this process:

The Pitman-Yor process

Draws from the Pitman-Yor process $G_1 \sim PY(d, \theta, G_0)$ with a discount parameter $0 \leq d < 1$, a strength parameter $\theta > -d$, and a base distribution G_0, are constructed using a Chinese restaurant process as follows:

$$X_{c_\cdot+1} | X_1, ..., X_{c_\cdot} \sim \sum_{k=1}^{t_\cdot} \frac{c_k - d}{\theta + c_\cdot} \delta_{y_k} + \frac{\theta + t_\cdot d}{\theta + c_\cdot} G_0$$

Where c_k denotes the number of X_is (tokens) assigned to y_k (a type) and $t_\cdot$ is the total number of y_ks drawn from G_0.

The hierarchical Pitman-Yor process is simply a Pitman-Yor process where the base distribution is itself a Pitman-Yor process.

The hierarchical Pitman-Yor process

Denoting a context of atoms $\mathbf{u}$ as $(w_{i-l}, ..., w_{i-1})$, the hierarchical Pitman-Yor process is defined using the above definition of the Pitman-Yor process by:

$$\begin{aligned} w_i &\sim G_{\mathbf{u}} \\ G_{\mathbf{u}} &\sim PY(d_{|\mathbf{u}|}, \theta_{|\mathbf{u}|}, G_{\pi(\mathbf{u})}) \\ &\dots \\ G_{(w_{i-1})} &\sim PY(d_1, \theta_1, G_{\emptyset}) \\ G_{\emptyset} &\sim PY(d_0, \theta_0, G_0) \end{aligned}$$

where $\pi(\mathbf{u}) = (w_{i-l+1}, ..., w_{i-1})$ is the suffix of $\mathbf{u}$, $|\mathbf{u}|$ denotes the length of context $\mathbf{u}$, and G_0 is a base distribution (usually uniform over all words).

UNIVERSITY OF CAMBRIDGE

Comparing this to interpolated Kneser-Ney discounting language model, we see that Kneser-Ney is simply a hierarchical Pitman-Yor process with parameter θ set to zero and a constraint of one table $t_{\mathbf{u}w} = 1$:

Interpolated Kneser-Ney discounting language model

$$P_{\mathbf{u}}(w) = \frac{\max(0, c_{\mathbf{u}w} - d_{|\mathbf{u}|})}{c_{\mathbf{u}\cdot}} + \frac{d_{|\mathbf{u}|} t_{\mathbf{u}\cdot}}{c_{\mathbf{u}\cdot}} P_{\pi(\mathbf{u})}(w)$$

where $c_{\mathbf{u}w}$ is the number of observations of the sequence $\mathbf{u}$ followed by the word w, $c_{\mathbf{u}\cdot}$ is the number of observations of the sequence $\mathbf{u}$ itself, and $t_{\mathbf{u}\cdot}$ is the number of unique words following the sequence $\mathbf{u}$.

The hierarchical Pitman-Yor process

$$P_{\mathbf{u}}(w) = \frac{c_{\mathbf{u}w} - d_{|\mathbf{u}|} t_{\mathbf{u}w}}{\theta + c_{\mathbf{u}\cdot}} + \frac{\theta + d_{|\mathbf{u}|} t_{\mathbf{u}\cdot}}{\theta + c_{\mathbf{u}\cdot}} P_{\pi(\mathbf{u})}(w)$$

- Modified Kneser-Ney uses different values of discounts for different counts

UNIVERSITY OF CAMBRIDGE

Bayesian Nonparametric approaches have been in use in NLP since the 90's!

Bayesian Nonparametric approaches

UNIVERSITY OF CAMBRIDGE

We can take advantage of the smoothing and interpolation with shorter contexts properties of the hierarchical Pitman-Yor (PY) process, and use it in word alignment as well.

Reminder: Model 1 generative story

$$P(F, A|E) = p(m|l) \times \prod_{i=1}^{m} p(a_i)p(f_i|e_{a_i})$$

Where $p(a_i) = \frac{1}{l+1}$ is uniform over all alignments and $p(f_i|e_{a_i}) \sim \mathit{Categorical}$.

- F and E are the input (source) and output (target) sentences of length J and I respectively,
- A is a vector of length J consisting of integer indexes into the target sentence – the alignment.

Re-formulating the model to use the hierarchical PY process instead of the categorical distributions, we get:

PY Model 1 generative story

$$
\begin{aligned}
a_i | m &\sim G_0^m \\
f_i | e_{a_i} &\sim H_{e_{a_i}} \\
H_{e_{a_i}} &\sim PY(H_\emptyset) \\
H_\emptyset &\sim PY(H_0)
\end{aligned}
$$

- f_i and a_i are the i'th foreign word and its alignment position,
- e_{a_i} is the English word corresponding to alignment position a_i,
- m is the lengths of the foreign sentence.

UNIVERSITY OF CAMBRIDGE

Following this approach, we can re-formulate the HMM alignment model as well to use the hierarchical PY process instead of the categorical distributions.

Reminder: HMM alignment model generative story

$$P(F,A|E) = p(m|l) \times \prod_{i=1}^{m} p(a_i|a_{i-1}, m) \times p(f_i|e_{a_i})$$

- f_i and a_i are the i'th foreign word and its alignment position,
- e_{a_i} is the English word corresponding to alignment position a_i,
- m and l are the lengths of the foreign and English sentences respectively.

UNIVERSITY OF CAMBRIDGE

We replace the categorical distribution for the transition $p(a_i|a_{i-1}, m)$ with a hierarchical PY process with a longer sequence of alignment positions in the conditional

PY HMM alignment model generative story

$$\begin{aligned} a_i|a_{i-1}, m &\sim G^m_{a_{i-1}} \\ G^m_{a_{i-1}} &\sim PY(G^m_{\emptyset}) \\ G^m_{\emptyset} &\sim PY(G^m_0) \end{aligned}$$

- Unique distribution for each foreign sentence length
- Condition the position on the previous alignment position, backing-off to the HMM's stationary distribution over alignment positions

Bayesian Nonparametric approaches

Unlike previous approaches that ran into difficulties extending models 3 and 4, we can extend them rather easily by just replacing the categorical distributions.

- The inference method that we use, Gibbs sampling, circumvents the intractable sum approximation of other inference methods
- The use of the hierarchical PY process allows us to incorporate phrasal dependencies into the distribution

UNIVERSITY OF CAMBRIDGE

Reminder: Models 3 and 4 generative story

$$P(F,A|E) = p(B_0|B_1,...,B_I) \times \prod_{i=1}^{I} p(B_i|B_{i-1},e_i) \times \prod_{i=0}^{I} \prod_{j \in B_i} p(f_j|e_i)$$

For model 3 the dependence on previous alignment sets is ignored and the probability $p(B_i|B_{i-1},e_i)$ is modelled as

$$p(B_i|B_{i-1},e_i) = p(\phi_i|e_i)\phi_i! \prod_{j \in B_i} p(j|i,m),$$

whereas in model 4 it is modelled using two HMMs:

$$\begin{aligned} p(B_i|B_{i-1},e_i) = & p(\phi_i|e_i) \times p_{=1}(B_{i,1} - \odot(B_{i-1})|\cdot) \\ & \times \prod_{k=2}^{\phi_i} p_{>1}(B_{i,k} - B_{i,k-1}|\cdot) \end{aligned}$$

Replacing the categorical priors with hierarchical PY process ones, we set the translation and fertility probabilities $p(\phi_i|e_i)\prod_{j\in B_i} p(f_j|e_i)$ using a common prior that generates translation sequences.

PY models 3 and 4 generative story

$$(f^1, \ldots, f^{\phi_i})|e_i \sim H_{e_i}$$

$$H_{e_i} \sim PY(H^{FT}_{e_i})$$

$$H^{FT}_{e_i}((f^1, \ldots, f^{\phi_i})) = H^F_{e_i}(\phi_i)\prod_j H^T_{(f^{j-1}, e_i)}(f^j)$$

$$H^F_{e_i} \sim PY(H^F_{\emptyset}) \qquad H^T_{(f^{j-1}, e_i)} \sim PY(H^T_{e_i})$$

$$H^F_{\emptyset} \sim PY(H^F_0) \qquad H^T_{e_i} \sim PY(H^T_{\emptyset})$$

$$H^T_{\emptyset} \sim PY(H^T_0)$$

- We used superscripts for the indexing of words which do not have to occur sequentially in the sentence

We generate sequences instead of individual words and fertilities, and fall-back onto these only in sparse cases.

Example

Aligning the English sentence "I don't speak French" to its French translation "Je ne parle pas français", the word "not" will generate the phrase ("ne", "pas"), which will later on be distorted into its place around the verb.

- The distortion probability for model 3, $p(j|i, m)$, is modelled as depending on the position of the source word i and its class
  - Interpolating for sparsity
  - The same way the HMM model backs-off to shorter sequences
- Similarly for the two HMMs in model 4.

How does this model compare to the EM trained models?

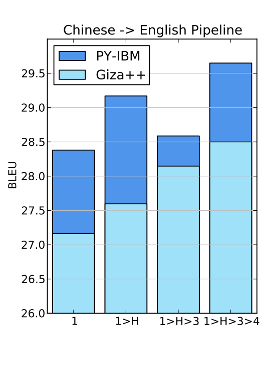


Figure: BLEU scores of pipelined Giza++ and pipelined PY-IBM translating from Chinese into English

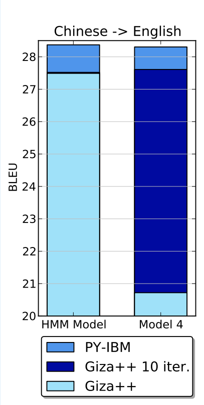


Figure: BLEU scores of Giza++'s and PY-IBM's HMM model and model 4 translating from Chinese into English

Limitations

- The use of Gibbs sampling for inference in this model is slow
  - On bi-corpora limited in size ($\sim$500K sentence pairs) the training takes 12 hours, compared to one hour for the EM model
  - More suitable for language pairs with high divergence – captures information that is otherwise lost

Outline

UNIVERSITY OF CAMBRIDGE

Arabic → English

بغداد 1-1 (افب) - ذكرت وكالة الانباء العراقية الرسمية ان نائب رئيس
مجلس قيادة الثورة في العراق عزة ابراهيم استقبل اليوم الاربعاء في بغداد
رئيس مجلس ادارة المركز السعودي لـ- تطوير الصادرات عبد الرحمن الزامل .

↓

?

Conclusions

Arabic → English

بغداد 1-1 (افب) - ذكرت وكالة الانباء العراقية الرسمية ان نائب رئيس
مجلس قيادة الثورة في العراق عزة ابراهيم استقبل اليوم الاربعاء في بغداد
رئيس مجلس ادارة المركز السعودي ل- تطوير الصادرات عبد الرحمن الزامل .

↓

Baghdad 1-1 (AFP) - official Iraqi news agency reported that vice-chairman of the revolution command council Izzat Ibrahim received in Iraq on Wednesday in Baghdad, board chairman of the Saudi center for developing exports Abdel Rahman Al-Zamil.

- Statistical machine translation works!

Chinese → English

加拿大与欧盟和澳洲一样 都在十一月二十八日关闭它们的大使馆,并在本周稍早重新开放。

↓

Canada and the EU and Australia have closed on 28 November at the same as the Chinese embassy in their earlier this week, and re-opening up.

- Statistical machine translation works ... sometimes!

Conclusions

UNIVERSITY OF CAMBRIDGE

- Statistical machine translation is a fully functional commercial technology
- Lots of linguistic challenges remain:
  - long distance reordering
  - complex morphology
  - underspecification
- Lots of theoretical and engineering challenges to be explored:
  - approximate search for intractable models
  - automatic learning of syntactic and semantic structures
  - efficiently dealing with massive quantities of data
- Lots of room for improvement with latest Bayesian research
- Many potential research projects!
- Real-world application for the evaluation of new techniques!

References 1:

- Daniel Jurafsky, James H. Martin, *An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Second edition.
- Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., and Mercer, R. L., *The mathematics of statistical machine translation: Parameter estimation*, Computational Linguistics, 19(2), 263–311 (1993).
- Stephan Vogel, Hermann Ney, Christoph Tillmann, *HMM-based word alignment in statistical translation*, Proceedings of the 16th conference on Computational linguistics, August 05-09, 1996, Copenhagen, Denmark.
- Franz Josef Och, Hermann Ney, *A Systematic Comparison of Various Statistical Alignment Models*, Computational Linguistics, v.29 n.1, p.19-51, March 2003.
- Coskun Mermer, Murat Saraclar *Bayesian Word Alignment for Statistical Machine Translation*, In Proceedings of ACL HLT, 2011.

References 2:

- Yee Whye Teh, *A hierarchical Bayesian language model based on Pitman-Yor processes*, Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, p.985-992, Sydney, Australia, July 17-18, 2006.
- Y. W. Teh, *A Bayesian Interpretation of Interpolated Kneser-Ney NUS School of Computing Technical Report TRA2/06*, Technical Report TRA2/06, School of Computing, National University of Singapore, 2006.
- Riley, Darcey and Gildea, Daniel, *Improving the IBM alignment models using variational Bayes*, Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2, ACL '12, 2012
- Gal, Yarin and Blunsom, Phil, *A Systematic Bayesian Treatment of the IBM Alignment Models*, Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, June, 2013

UNIVERSITY OF CAMBRIDGE

Questions?

Human Evaluation

There are two dominant approaches to the subjective evaluation of automatic translation:

- Scoring (1–5) individual sentences based on:
  - *adequacy*: does it preserve the meaning?
  - *fluency*: is it real language?
- Comparing sentences produced by two different systems
  - *binary comparison*: is the sentence output from system A better than that from system B?
  - *ranking*: rank the outputs of X systems?

Eliciting such evaluations is slow, expensive, and because human judges often don't agree, unreliable. However human evaluation remains the gold standard for comparing translation models.

Human Evaluation

UNIVERSITY OF CAMBRIDGE

How would you rank this translation?

Source: 欧盟 办事处 与 澳洲 大使馆 在 同 一 建筑 内

Candidate: the chinese embassy in australia and the eu representative office in the same building

Reference Translations:

1. the eu office and the australian embassy are housed in the same building

UNIVERSITY OF CAMBRIDGE

Ngram overlap metrics:

Source: 欧盟 办事处 与 澳洲 大使馆 在 同 一 建筑 内

Candidate: the chinese embassy in australia and the eu representative office in the same building

Reference Translations:

1. the eu office and the australian embassy are housed in the same building
2. the european union office is in the same building as the australian embassy
3. the european union 's office and the australian embassy are both located in the same building
4. the eu 's mission is in the same building with the australian embassy

UNIVERSITY OF CAMBRIDGE

Ngram overlap metrics: 1-gram precision $p_1 = \frac{11}{14}$

Source: 欧盟 办事处 与 澳洲 大使馆 在 同 一 建筑 内

Candidate: the chinese embassy in australia and the eu representative office in the same building

Reference Translations:

1. the eu office and the australian embassy are housed in the same building
2. the european union office is in the same building as the australian embassy
3. the european union 's office and the australian embassy are both located in the same building
4. the eu 's mission is in the same building with the australian embassy

UNIVERSITY OF CAMBRIDGE

Ngram overlap metrics: 2-gram precision $p_2 = \frac{5}{13}$

Source: 欧盟 办事处 与 澳洲 大使馆 在 同 一 建筑 内

Candidate: the chinese embassy in australia and the eu representative office in the same building

Reference Translations:

1. the eu office and the australian embassy are housed in the same building
2. the european union office is in the same building as the australian embassy
3. the european union 's office and the australian embassy are both located in the same building
4. the eu 's mission is in the same building with the australian embassy

UNIVERSITY OF CAMBRIDGE

Ngram overlap metrics: 3-gram precision $p_3 = \frac{2}{12}$

Source: 欧盟 办事处 与 澳洲 大使馆 在 同 一 建筑 内

Candidate: the chinese embassy in australia and the eu representative office in the same building

Reference Translations:

1. the eu office and the australian embassy are housed in the same building
2. the european union office is in the same building as the australian embassy
3. the european union 's office and the australian embassy are both located in the same building
4. the eu 's mission is in the same building with the australian embassy

UNIVERSITY OF CAMBRIDGE

Ngram overlap metrics: 4-gram precision $p_4 = \frac{1}{11}$

Source: 欧盟 办事处 与 澳洲 大使馆 在 同 一 建筑 内

Candidate: the chinese embassy in australia and the eu representative office in the same building

Reference Translations:

1. the eu office and the australian embassy are housed in the same building
2. the european union office is in the same building as the australian embassy
3. the european union 's office and the australian embassy are both located in the same building
4. the eu 's mission is in the same building with the australian embassy

UNIVERSITY OF CAMBRIDGE

Numerous automatic evaluation functions have been proposed, however the dominant metric is *BLEU*:

BLEU

$$BLEU_n = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$BP = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{R'}{C'}\right) & \text{if } c <= r \end{cases}$$

- *BP* is the *Brevity Penalty*, w_n is the ngram length weights (usually $\frac{1}{n}$), p_n is precision of ngram predictions, R' is the total length of all references and C' is the sum of the best matching candidates.
- statistics are calculate over the whole *document*, i.e. all the sentences.

UNIVERSITY OF CAMBRIDGE

Questions?

We can take advantage of the smoothing and interpolation with shorter contexts properties of the hierarchical Pitman-Yor (PY) process, and use it in word alignment as well.

Reminder: Model 1 generative story

$$P(F, A|E) = p(m|l) \times \prod_{i=1}^{m} p(a_i)p(f_i|e_{a_i})$$

Where $p(a_i) = \frac{1}{l+1}$ is uniform over all alignments and $p(f_i|e_{a_i}) \sim \mathit{Categorical}$.

Re-formulating the model to use the hierarchical PY process instead of the categorical distributions, we get:

PY Model 1 generative story

$$
\begin{aligned}
a_i | m &\sim G_0^m \\
f_i | e_{a_i} &\sim H_{e_{a_i}} \\
H_{e_{a_i}} &\sim PY(H_\emptyset) \\
H_\emptyset &\sim PY(H_0)
\end{aligned}
$$

Bayesian Nonparametric approaches

Following this approach, we can re-formulate the HMM alignment model as well to use the hierarchical PY process instead of the categorical distributions.

Reminder: HMM alignment model generative story

$$P(F,A|E) =$$
$$p(m|l) \times \prod_{i=1}^{m} p(a_i|a_{i-1}, m) \times p(f_i|e_{a_i})$$

Bayesian Nonparametric approaches

We replace the categorical distribution for the transition $p(a_i|a_{i-1}, m)$ with a hierarchical PY process with a longer sequence of alignment positions in the conditional

PY HMM alignment model generative story

$$
\begin{aligned}
a_i|a_{i-1}, m &\sim G^m_{a_{i-1}} \\
G^m_{a_{i-1}} &\sim PY(G^m_\emptyset) \\
G^m_\emptyset &\sim PY(G^m_0)
\end{aligned}
$$

- Unique distribution for each foreign sentence length
- Condition the position on the previous alignment position, backing-off to the HMM's stationary distribution over alignment positions

Bayesian Nonparametric approaches

Unlike previous approaches that ran into difficulties extending models 3 and 4, we can extend them rather easily by just replacing the categorical distributions.

- The inference method that we use, Gibbs sampling, circumvents the intractable sum approximation of other inference methods
- The use of the hierarchical PY process allows us to incorporate phrasal dependencies into the distribution

Reminder: Models 3 and 4 generative story

$$P(F, A|E) = p(B_0|B_1, ..., B_I) \times \prod_{i=1}^{I} p(B_i|B_{i-1}, e_i) \times \prod_{i=0}^{I} \prod_{j \in B_i} p(f_j|e_i)$$

UNIVERSITY OF CAMBRIDGE

Reminder: Models 3 and 4 generative story – cont.

For model 3 the dependence on previous alignment sets is ignored and the probability $p(B_i|B_{i-1}, e_i)$ is modelled as

$$p(B_i|B_{i-1}, e_i) = p(\phi_i|e_i)\phi_i! \prod_{j \in B_i} p(j|i, m),$$

whereas in model 4 it is modelled using two HMMs:

$$\begin{aligned} p(B_i|B_{i-1}, e_i) =& p(\phi_i|e_i) \times p_{=1}(B_{i,1} - \odot(B_{i-1})|\cdot) \\ & \times \prod_{k=2}^{\phi_i} p_{>1}(B_{i,k} - B_{i,k-1}|\cdot) \end{aligned}$$

Reminder: Models 3 and 4 generative story – cont.

For both these models the spurious word generation is controlled by a binomial distribution:

$$p(B_0|B_1,...,B_l) = \binom{m-\phi_0}{\phi_0}(1-p_0)^{m-2\phi_0}p_1^{\phi_0}\frac{1}{\phi_0!}$$

for some parameters p_0 and p_1.

Bayesian Nonparametric approaches

Replacing the categorical priors with hierarchical PY process ones, we set the translation and fertility probabilities $p(\phi_i|e_i)\prod_{j\in B_i} p(f_j|e_i)$ using a common prior that generates translation sequences.

PY models 3 and 4 generative story

$$(f^{1}, \dots, f^{\phi_i})|e_i \sim H_{e_i}$$

$$H_{e_i} \sim PY(H^{FT}_{e_i})$$

$$H^{FT}_{e_i}((f^{1}, \dots, f^{\phi_i})) = H^{F}_{e_i}(\phi_i)\prod_{j} H^{T}_{(f^{j-1},e_i)}(f^{j})$$

$$H^{F}_{e_i} \sim PY(H^{F}_{\emptyset}) \qquad H^{T}_{(f^{j-1},e_i)} \sim PY(H^{T}_{e_i})$$

$$H^{F}_{\emptyset} \sim PY(H^{F}_{0}) \qquad H^{T}_{e_i} \sim PY(H^{T}_{\emptyset})$$

$$H^{T}_{\emptyset} \sim PY(H^{T}_{0})$$

- We used superscripts for the indexing of words which do not have

We generate sequences instead of individual words and fertilities, and fall-back onto these only in sparse cases.

Example

Aligning the English sentence "I don't speak French" to its French translation "Je ne parle pas français", the word "not" will generate the phrase ("ne", "pas"), which will later on be distorted into its place around the verb.

UNIVERSITY OF CAMBRIDGE

The distortion probability for model 3, $p(j|i, m)$, is modelled simply as depending on the position of the source word i and its class:

PY models 3 and 4 generative story – cont.

$$\begin{aligned} j|(C(e_i), i), m &\sim G^m_{(C(e_i),i)} \\ G^m_{(C(e_i),i)} &\sim PY(G^m_i) \\ G^m_i &\sim PY(G^m_\emptyset) \\ G^m_\emptyset &\sim PY(G^m_0) \end{aligned}$$

where we back-off to the source word position and then to the frequencies of the alignment positions.

Bayesian Nonparametric approaches

Distortion probability for IBM model 4

- First probability distribution $p_{=1}$ controls the head distortion

PY models 3 and 4 generative story – cont.

$$
\begin{aligned}
B_{i,1} - \odot(B_{i-1}) \mid (C(e_i), C(f_{B_{i,1}})), m & \\
&\sim G^{m}_{(C(e_i),C(f_{B_{i,1}}))} \\
G^{m}_{(C(e_i),C(f_{B_{i,1}}))} &\sim PY(G^{m}_{C(f_{B_{i,1}})}) \\
G^{m}_{C(f_{B_{i,1}})} &\sim PY(G^{m}_{\emptyset}) \\
G^{m}_{\emptyset} &\sim PY(G^{m}_{0})
\end{aligned}
$$

- Second probability distribution $p_{>1}$ controls the distortion within the set of words

PY models 3 and 4 generative story – cont.

$$B_{i,j} - B_{i,j-1} | C(f_{B_{i,j}}), m \sim H^m_{C(f_{B_{i,j}})}$$
$$H^m_{C(f_{B_{i,j}})} \sim PY(H^m_{\emptyset})$$
$$H^m_{\emptyset} \sim PY(H^m_0)$$

Again we model the jump size as depending on the word class for the proposed foreign word, backing-off to the relative jump frequencies.

UNIVERSITY OF CAMBRIDGE

Fertility and translation of NULL words

- Follows the idea of the original model, where the number of spurious words is determined by a binomial distribution created from a set of Bernoulli experiments, each one performed after the translation of a non-spurious word
- We use an indicator function I to signal whether a spurious word was generated after a non-spurious word ($I = 1$) or not ($I = 0$)

PY models 3 and 4 generative story – cont.

$$
\begin{aligned}
I = 0, 1 | l &\sim H_l^{NF} \\
H_l^{NF} &\sim PY(H_\emptyset^{NF}) \\
H_\emptyset^{NF} &\sim PY(H_0^{NF})
\end{aligned}
\qquad
\begin{aligned}
f_i &\sim H_\emptyset^{NT} \\
H_\emptyset^{NT} &\sim PY(H_0^{NT})
\end{aligned}
$$

Questions?

Word Alignment (IBM Model 1)

A simple generative model for $p(\mathbf{s}|\mathbf{t})$ is derived by introducing a latent variable $\mathbf{a}$ into the conditional probability:

$$p(\mathbf{s}, \mathbf{a}|\mathbf{t}) = \frac{p(J|I)}{(I+1)^J} \prod_{j=1}^{J} p(s_j|t_{a_j}),$$

where:

- $\mathbf{s}$ and $\mathbf{t}$ are the input (source) and output (target) sentences of length J and I respectively,
- $\mathbf{a}$ is a vector of length J consisting of integer indexes into the target sentence, known as the alignment,
- $p(J|I)$ is not important for training the model and we'll treat it as a constant ϵ.

To learn this model the EM algorithm is used to find the MLE values for the parameters $p(s_j|t_{a_j})$.

EM for Word Alignment (IBM Model 1)

UNIVERSITY OF CAMBRIDGE

To derive an EM update for this model we need to calculate the expected values for the alignment vectors for each sentence. The conditional probability of an alignment is:

$$p(\mathbf{a}|\mathbf{s},\mathbf{t}) = \frac{p(\mathbf{s},\mathbf{a}|\mathbf{t})}{p(\mathbf{s}|\mathbf{t})}$$

Marginalising out $\mathbf{a}$ in $p(\mathbf{s},\mathbf{a}|\mathbf{t})$ gives the required denominator:

$$\begin{aligned} p(\mathbf{s}|\mathbf{t}) &= \sum_{\mathbf{a}} p(\mathbf{s},\mathbf{a}|\mathbf{t}), \\ &= \sum_{a_1=0}^{I} \sum_{a_2=0}^{I} \cdots \sum_{a_J=0}^{I} p(\mathbf{s},\mathbf{a}|\mathbf{t}), \\ &= \frac{\epsilon}{(I+1)^J} \sum_{a_1=0}^{I} \sum_{a_2=0}^{I} \cdots \sum_{a_J=0}^{I} \prod_{j=1}^{J} p(s_j|t_{a_j}). \end{aligned}$$

Rather conveniently we can swap the sum and product to get an equation that is tractable to compute:

$$
\begin{aligned}
p(\mathbf{s}|\mathbf{t}) &= \frac{\epsilon}{(I+1)^J} \sum_{a_1=0}^{I} \sum_{a_2=0}^{I} \cdots \sum_{a_J=0}^{I} \prod_{j=1}^{J} p(s_j|t_{a_j}) \\
&= \frac{\epsilon}{(I+1)^J} \prod_{j=1}^{J} \sum_{i=0}^{I} p(s_j|t_i).
\end{aligned}
$$

UNIVERSITY OF CAMBRIDGE

Now we can state the conditional probabilities for the alignments:

$$
\begin{aligned}
p(\mathbf{a}|\mathbf{s},\mathbf{t}) &= \frac{p(\mathbf{s},\mathbf{a}|\mathbf{t})}{p(\mathbf{s}|\mathbf{t})}, \\
&= \frac{\frac{\epsilon}{(I+1)^J}\prod_{j=1}^{J} p(s_j|t_{a_j})}{\frac{\epsilon}{(I+1)^J}\prod_{j=1}^{J}\sum_{i=0}^{I} p(s_j|t_i)}, \\
&= \prod_{j=1}^{J}\frac{p(s_j|t_{a_j})}{\sum_{i=0}^{I} p(s_j|t_i)}
\end{aligned}
$$

UNIVERSITY OF CAMBRIDGE

The next step is to derive the expected counts $c(s|t, \mathbf{s}, \mathbf{t})$ for a single pair on sentences of a source word s aligning with a target word t:

$$
\begin{aligned}
c(s|t, \mathbf{s}, \mathbf{t}) &= \sum_{\mathbf{a}} p(\mathbf{a}|\mathbf{s}, \mathbf{t}) \sum_{j=1}^{J} \delta(s, s_j)\delta(t, t_{a_j}) \\
&= \frac{p(s|t)}{\sum_{i=0}^{I} p(s|t_i)} \sum_{j=1}^{J} \delta(s, s_j) \sum_{i=1}^{I} \delta(t, t_i)
\end{aligned}
$$

where we've used a similar trick to that used earlier to rearrange the sums. The result is that we can calculate the counts in $\mathcal{O}(J \times I)$ rather than $\mathcal{O}(I+1)^J$.

EM for Word Alignment (IBM Model 1)

Finally by collecting the counts for all sentence pairs in our training corpus $(\mathbf{s}, \mathbf{t})$ and normalising we can derive the EM update for the translation probabilities $p(s|t)$:

$$p^{i+1}(s|t) = \frac{\sum_{\mathbf{s},\mathbf{t}} c^i(s|t, \mathbf{s}, \mathbf{t})}{\sum_t \sum_{\mathbf{s},\mathbf{t}} c^i(s|t, \mathbf{s}, \mathbf{t})}.$$

EM for Word Alignment (IBM Model 1)

Algorithm outline:

1. Initialise the translation probabilities $p(s|t)$ to uniform,
2. **E Step:** For each pair of sentences in the training corpus, calculate $c^i(s|t, \mathbf{s}, \mathbf{t})$, keeping a running sum of $c^i(s|t)$ and $\sum_t p(s|t)$,
3. **M Step:** Calculate the new probabilities $p(s|t)$ using the normalised counts,
4. Repeat from 2 until the log likelihood of the data $(\sum_{\mathbf{s},\mathbf{t}} \log p(\mathbf{s}|\mathbf{t}))$ stops increasing
(up to a small tolerance).

Word Alignment (IBM Model 1)

Limitations of this simple word alignment model:

- The structure of sentences is not modelled, words align independently of each other,
- The position of words with a sentence is not modelled, obviously words near the start of the source sentence are more likely to align to words near the start of the target sentence,
- The alignment is asymmetric, a target word may align to multiple source words, but a source word may only align to a single target,
- and many others ...

These limitations mean that this model does not work well as a translation model on it's own, however it is currently used as the first step in learning more complicated models by online translation providers such as Google and Microsoft.

Questions?