Phrase-based SMT

Sophie Arnoult

Universiteit van Amsterdam

April 12, 2016

Content

The Noisy-Channel approach

Bayes rule

$$P(E|F) = \frac{P(E)P(F|E)}{P(F)}$$

Inference

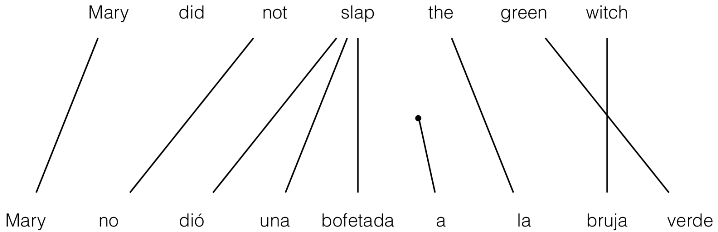
$$\hat{E} = \arg\max_{E} P(E)P(F|E)$$

Estimation

- $P(E)$ n-gram LM
- $P(F|E)$ ...

The IBM models

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



Models 1 and 2

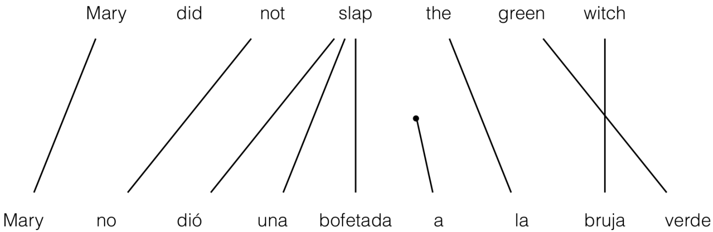
$$
\begin{aligned}
P(F, A|E) &= P(m|E) \prod_{j=1}^{m} P(a_j, f_j | a_1^{j-1}, f_1^{j-1}, m, E) \\
&= P(m|E) \prod_{j=1}^{m} P(a_j | a_1^{j-1}, f_1^{j-1}, m, E) P(f_j | a_1^{j}, f_1^{j-1}, m, E)
\end{aligned}
$$

Models 1 and 2

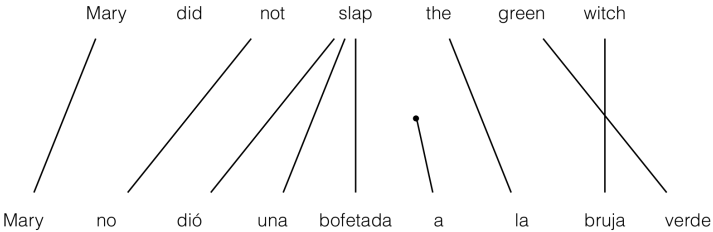
$$P(F,A|E) = P(m|E)\prod_{j=1}^{m} P(a_j, f_j | a_1^{j-1}, f_1^{j-1}, m, E)$$
$$= P(m|E)\prod_{j=1}^{m} P(a_j | a_1^{j-1}, f_1^{j-1}, m, E) P(f_j | a_1^{j}, f_1^{j-1}, m, E)$$

- lexical translation $P(f_j | a_1^j, f_1^{j-1}, m, E) = t(f_j | e_i)$
- alignment $P(a_j | \ldots)$
  - IBM1: $\sim unif(l+1)$
  - IBM2: $= a(i|j,m,l)$
  - HMM: $= a(i|a_{j-1}, l)$

Decoding with models 1 & 2?



Decoding with models 1 & 2?



how to explain insertions on the English side?

Modelling word fertility

- *fertility*: number of words generated by an English words
- Generative story
  - choose fertility for e_i
  - choose French words generated for each e_i
  - reorder French words

Modelling word fertility

- *fertility*: number of words generated by an English words
- Generative story
  - choose fertility for e_i
  - choose French words generated for each e_i
  - reorder French words
- parameters: fertility, translation, distortion, null-word

Modelling word fertility

- *fertility*: number of words generated by an English words
- Generative story
  - choose fertility for e_i
  - choose French words generated for each e_i
  - reorder French words
- parameters: fertility, translation, distortion, null-word
- inference is intractable:
  E step in neighbourhood of Viterbi alignment

Generative story

Mary	did	not	slap		the	green	witch

Generative story

Mary	did	not		slap			the	green	witch
Mary	~~did~~	not	slap	slap	slap		the	green	witch

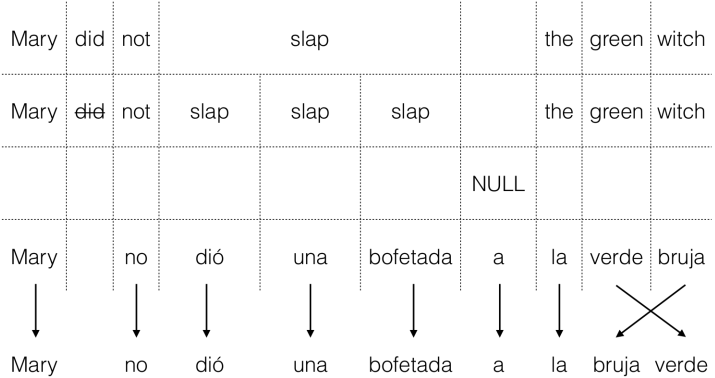
Generative story

Mary	did	not		slap			the	green	witch
Mary	~~did~~	not	slap	slap	slap		the	green	witch
						NULL			

Generative story

Mary	did	not	slap				the	green	witch
Mary	~~did~~	not	slap	slap	slap		the	green	witch
						NULL			
Mary		no	dió	una	bofetada	a	la	verde	bruja

Generative story



Conclusion

	Inference	Generation
1	Exact	Local search (and distortion limit)
2	Exact	
≥3	Approximate	

- IBM models 1 and 2 are too weak for decoding
- decoding is NP-complete (for phrase-based models too)
- asymmetry is unsatisfactory from linguistic perspective

From word-based to phrase-based SMT

Capturing non-compositional translation equivalents

- multi-word expressions
  - (Fr) “est-ce que” $\leftrightarrow$ “do/did”
  - “kick the bucket” $\leftrightarrow$ “die”

From word-based to phrase-based SMT

Capturing non-compositional translation equivalents

- multi-word expressions
  - (Fr) “est-ce que” $\leftrightarrow$ “do/did”
  - “kick the bucket” $\leftrightarrow$ “die”
- morphology / inflection
  - very limited in English
  - verb inflection and noun agreement in Romance languages

From word-based to phrase-based SMT

Capturing non-compositional translation equivalents

- multi-word expressions
  - (Fr) "est-ce que" $\leftrightarrow$ "do/did"
  - "kick the bucket" $\leftrightarrow$ "die"
- morphology / inflection
  - very limited in English
  - verb inflection and noun agreement in Romance languages
  - "est-ce que tu voulais" $\leftrightarrow$ "did you want"
    (*? you want-*Past*-you* $\leftrightarrow$ *?-*Past *you want*)
  - "tu as gagné / gagnais " $\not\leftrightarrow$ "you won / have won" (aspect)

From word-based to phrase-based SMT

Capturing non-compositional translation equivalents

- multi-word expressions
  - (Fr) “est-ce que” $\leftrightarrow$ “do/did”
  - “kick the bucket” $\leftrightarrow$ “die”
- morphology / inflection
  - very limited in English
  - verb inflection and noun agreement in Romance languages
  - “est-ce que tu voulais” $\leftrightarrow$ “did you want”
    (*? you want-Past-you* $\leftrightarrow$ *?-Past you want*)
  - “tu as gagné / gagnais ” $\not\leftrightarrow$ “you won / have won” (aspect)
- local reorderings
  - “un homme grand” $\leftrightarrow$ “a tall man”
  - “un grand homme”$\leftrightarrow$ “a great man”

Example

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

Generative story

A new hidden variable: segmentation S

One possible story

$$\begin{aligned} P(F|E) &= \sum_S \sum_A P(S, A, F|E) \\ &= \sum_S \sum_A P(S|E) \times P(A|S, E) \times P(F|A, S, E) \end{aligned}$$

Generative story

A new hidden variable: segmentation S

One possible story

$$\begin{aligned} P(F|E) &= \sum_S \sum_A P(S, A, F|E) \\ &= \sum_S \sum_A P(S|E) \times P(A|S, E) \times P(F|A, S, E) \end{aligned}$$

I_1 have_2 black_3 eyes_4 input

Generative story

A new hidden variable: segmentation S

One possible story

$$P(F|E) = \sum_S \sum_A P(S, A, F|E)$$
$$= \sum_S \sum_A P(S|E) \times P(A|S, E) \times P(F|A, S, E)$$

I_1 have_2 black_3 eyes_4	input
[I_1 have_2] [black_3] [eyes_4]	segmentation

Generative story

A new hidden variable: segmentation S

One possible story

$$P(F|E) = \sum_S \sum_A P(S, A, F|E)$$
$$= \sum_S \sum_A P(S|E) \times P(A|S, E) \times P(F|A, S, E)$$

I_1 have_2 black_3 eyes_4	input
[I_1 have_2] [black_3] [eyes_4]	segmentation
[I_1 have_2]$_1$ [eyes_4]$_3$ [black_3]$_2$	ordering

Generative story

A new hidden variable: segmentation S

One possible story

$$P(F|E) = \sum_S \sum_A P(S, A, F|E)$$
$$= \sum_S \sum_A P(S|E) \times P(A|S, E) \times P(F|A, S, E)$$

I_1 have_2 black_3 eyes_4	input
[I_1 have_2] [black_3] [eyes_4]	segmentation
[I_1 have_2]$_1$ [eyes_4]$_3$ [black_3]$_2$	ordering
[J' ai]$_1$ [les yeux]$_3$ [noirs]$_2$	translation

(MLE) inference in phrase-based models

- [Marcu and Wong, 2002]
    - hidden segmentation and alignment
    - uniform segmentation, infer distortion and translation probabilities

(MLE) inference in phrase-based models

- [Marcu and Wong, 2002]
  - hidden segmentation and alignment
  - uniform segmentation, infer distortion and translation probabilities
  - approximate inference, overfitting

(MLE) inference in phrase-based models

- [Marcu and Wong, 2002]
  - hidden segmentation and alignment
  - uniform segmentation, infer distortion and translation probabilities
  - approximate inference, overfitting
- [DeNero et al., 2006]
  - comparable to [Marcu and Wong, 2002], with observed (word) alignments

(MLE) inference in phrase-based models

- [Marcu and Wong, 2002]
  - hidden segmentation and alignment
  - uniform segmentation, infer distortion and translation probabilities
  - approximate inference, overfitting
- [DeNero et al., 2006]
  - comparable to [Marcu and Wong, 2002], with observed (word) alignments
  - approximate inference, overfitting

(MLE) inference in phrase-based models

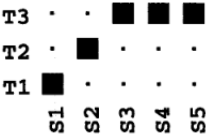
- [Marcu and Wong, 2002]
    - hidden segmentation and alignment
    - uniform segmentation, infer distortion and translation probabilities
    - approximate inference, overfitting
- [DeNero et al., 2006]
    - comparable to [Marcu and Wong, 2002], with observed (word) alignments
    - approximate inference, overfitting
- [Koehn et al., 2003]
    - observed (word) alignment and phrase pairs (not segmentations!)
    - parametric distortion and heuristic translation estimates

(MLE) inference in phrase-based models

- [Marcu and Wong, 2002]
  - hidden segmentation and alignment
  - uniform segmentation, infer distortion and translation probabilities
  - approximate inference, overfitting
- [DeNero et al., 2006]
  - comparable to [Marcu and Wong, 2002], with observed (word) alignments
  - approximate inference, overfitting
- [Koehn et al., 2003]
  - observed (word) alignment and phrase pairs (not segmentations!)
  - parametric distortion and heuristic translation estimates
- [Mylonakis and Sima'an, 2008]
  - observed (word) alignment and phrase pairs
  - ITG-based segmentation, infer translation probabilities

The Alignment-Template Approach

[Och and Ney, 2000] laid out the fundations for [Koehn et al., 2003]




```
T1:  zwei, drei, vier, fünf, ...
T2:  Uhr
T3:  vormittags, nachmittags, abends, ...

S1:  two, three, four, five, ...
S2:  o'clock
S3:  in
S4:  the
S5:  morning, evening, afternoon, ...
```

Alignment template: (class) phrase pair & internal alignment

Model

$$\text{score}(E, S, A|F) = \theta^\top h(F, E, A, S)$$

Model

$$\text{score}(E, S, A|F) = \theta^\top h(F, E, A, S)$$

features include

- language model
- alignment (distortion)
- translation

Model

$$\text{score}(E, S, A|F) = \theta^\top h(F, E, A, S)$$

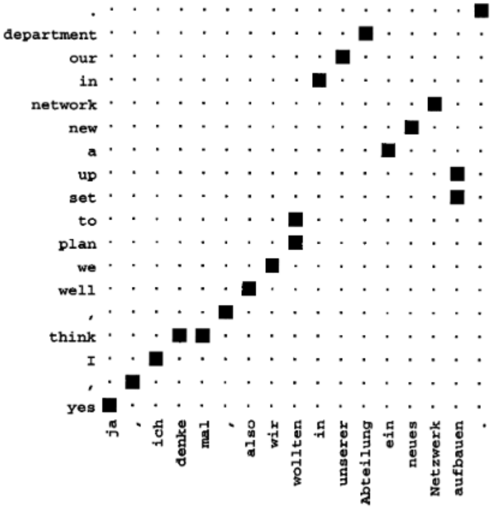
features include

- language model
- alignment (distortion)
- translation

independence assumptions

- $h_A(F, E, A, S) = \log \prod_k p(a_k|F, E, A, S)$
- $h_F(F, E, A, S) = \log \prod_k p(\bar{f}_k|F, E, A, S)$

Alignment symmetrization



Alignment consistency

Let $(\bar{f}, \bar{e})$ be a phrase pair
Let A be an alignment matrix

Alignment consistency

Let $(\bar{f}, \bar{e})$ be a phrase pair
Let A be an alignment matrix

$(\bar{f}, \bar{e})$ consistent with A if, and only if:

Alignment consistency

Let $(\bar{f}, \bar{e})$ be a phrase pair
Let A be an alignment matrix

$(\bar{f}, \bar{e})$ consistent with A if, and only if:

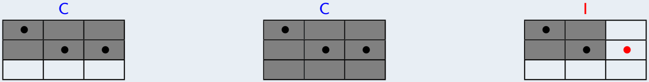
- Words in $\bar{f}$, if aligned, align only with words in $\bar{e}$

Alignment consistency

Let $(\bar{f}, \bar{e})$ be a phrase pair
Let A be an alignment matrix

$(\bar{f}, \bar{e})$ consistent with A if, and only if:

- Words in $\bar{f}$, if aligned, align only with words in $\bar{e}$



Alignment consistency

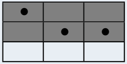
Let $(\bar{f}, \bar{e})$ be a phrase pair

Let A be an alignment matrix

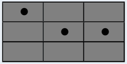
$(\bar{f}, \bar{e})$ consistent with A if, and only if:

- Words in $\bar{f}$, if aligned, align only with words in $\bar{e}$

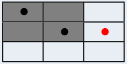
C



C



I



- Words in $\bar{e}$, if aligned, align only with words in $\bar{f}$

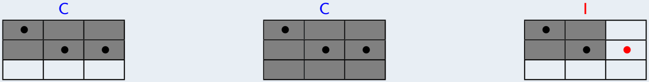
Alignment consistency

Let $(\bar{f}, \bar{e})$ be a phrase pair

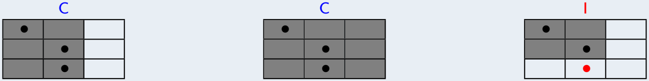
Let A be an alignment matrix

$(\bar{f}, \bar{e})$ consistent with A if, and only if:

- Words in $\bar{f}$, if aligned, align only with words in $\bar{e}$



- Words in $\bar{e}$, if aligned, align only with words in $\bar{f}$



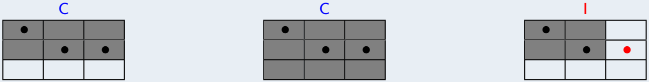
Alignment consistency

Let $(\bar{f}, \bar{e})$ be a phrase pair

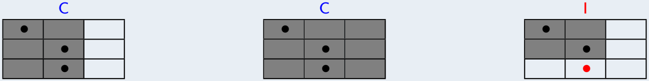
Let A be an alignment matrix

$(\bar{f}, \bar{e})$ consistent with A if, and only if:

- Words in $\bar{f}$, if aligned, align only with words in $\bar{e}$



- Words in $\bar{e}$, if aligned, align only with words in $\bar{f}$



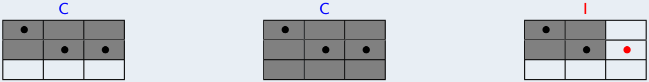
- $(\bar{f}, \bar{e})$ must contain at least one alignment point

Alignment consistency

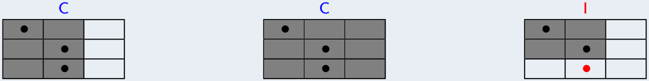
Let $(\bar{f}, \bar{e})$ be a phrase pair
Let A be an alignment matrix

$(\bar{f}, \bar{e})$ consistent with A if, and only if:

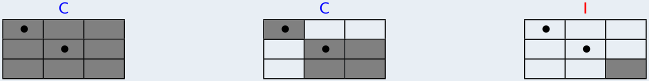
- Words in $\bar{f}$, if aligned, align only with words in $\bar{e}$



- Words in $\bar{e}$, if aligned, align only with words in $\bar{f}$



- $(\bar{f}, \bar{e})$ must contain at least one alignment point



Phrase extraction

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

Phrase extraction

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an "observed" phrase pair

Phrase extraction

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an "observed" phrase pair

Phrase extraction

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an "observed" phrase pair

Phrase extraction

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an "observed" phrase pair

Phrase extraction

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an "observed" phrase pair

Phrase extraction

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an "observed" phrase pair

Phrase extraction

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an "observed" phrase pair

Phrase extraction

		I	have	black	eyes
1	J'				
2	ai				
3	les				
4	yeux				
5	noirs				

- multiple derivations can explain an "observed" phrase pair
- we extract all of them once, irrespective of derivation

Translation estimates

Number of times a (consistent) phrase pair is "observed"

$$c(\bar{f}, \bar{e})$$

Relative frequency counting

$$\phi(\bar{f}|\bar{e}) = \frac{c(\bar{f}, \bar{e})}{\sum_{\bar{f}'} c(\bar{f}', \bar{e})}$$

Features

- language model
- forward translation probability $P(F|E)$
- backward translation probability $P(E|F)$
- forward and backward lexical smoothing
- word penalty
- phrase penalty
- distance-based reordering model
- lexical reordering model

Distance-based reordering

- exponential $\delta(d_k) = \alpha^{d_k}, \alpha < 1$
- $d_k = |\text{start}_k - \text{end}_{k-1} - 1|$

Distance-based reordering

- exponential $\delta(d_k) = \alpha^{d_k}, \alpha < 1$
- $d_k = |\text{start}_k - \text{end}_{k-1} - 1|$

<table>
<tr><td></td><td></td><td>I</td><td>have</td><td>black</td><td>eyes</td></tr>
<tr><td>1</td><td>J'</td><td colspan="2" rowspan="2">1</td><td></td><td></td></tr>
<tr><td>2</td><td>ai</td><td></td><td></td></tr>
<tr><td>3</td><td>les</td><td></td><td></td><td></td><td rowspan="2">3</td></tr>
<tr><td>4</td><td>yeux</td><td></td><td></td><td></td></tr>
<tr><td>5</td><td>noirs</td><td></td><td></td><td>2</td><td></td></tr>
</table>

- $\bar{f}_1 =$ J' ai
- $\bar{e}_1 =$ I have
- start$_1 = 1$
- end$_1 = 2$

- $\bar{f}_2 =$ noirs
- $\bar{e}_2 =$ black
- start$_2 = 5$
- end$_2 = 5$

- $\bar{f}_3 =$ les yeux
- $\bar{e}_3 =$ eyes
- start$_3 = 3$
- end$_3 = 4$

Conclusion

- generative modelling requires approximations
- overfitting in fragment models (DOP)
- [Koehn et al., 2003] ignore segmentation:
  good feature choice in discriminative model
- reordering remains an issue

Decoding

Disambiguation problem

$$\hat{E} = \arg\max_E P(E)P(F|E) = \arg\max_E P(E) \sum_A P(F, A|E)$$

NP-complete [Sima'an, 2002]

Decoding

Disambiguation problem

$$\begin{aligned}\hat{E} &= \arg\max_{E} P(E)P(F|E) \\ &= \arg\max_{E} P(E)\sum_{A} P(F,A|E)\end{aligned}$$

NP-complete [Sima'an, 2002]

Viterbi approximation

$$\hat{E} \approx \arg\max_{E,A} P(E)P(F,A|E)$$

Viterbi decoding

The alignment space (or space of *derivations*)

- $O(2^n)$ segmentations
- $O(n!)$ permutations
- $O(t^n)$ substitutions

Viterbi decoding

The alignment space (or space of *derivations*)

- $O(2^n)$ segmentations
- $O(n!)$ permutations
- $O(t^n)$ substitutions

Packed representation using finite-state transducers

$$O(n^2 \times 2^n \times t)$$

NP-complete (TSP) [Knight, 1999, Zaslavskiy et al., 2009]

Complete model

$$P(E)P(F,S|E) = \prod_{j=1}^{|E|} \psi(e_j|e_{j-n+1}^{j-1}) \prod_{i=1}^{|S|} \phi(\bar{f}_i|\bar{e}_i)\delta(\mathsf{start}_i - \mathsf{end}_{i-1} - 1)$$

Approximations:

- distortion limit d: $2^n \rightarrow 2^d$
- maximum phrase length m: $n^2 \rightarrow n \times m$

- alignment space $O(2^d \times n \times m \times t)$
- weighted derivations $O(2^d \times n \times m \times t \times |\Delta|^{k-1})$
  where $P(E)$ is a k-gram LM components over Δ^*
  and $|\Delta| \propto t \times n$

Complete model

$$P(E)P(F,S|E) = \prod_{j=1}^{|E|} \psi(e_j|e_{j-n+1}^{j-1}) \prod_{i=1}^{|S|} \phi(\bar{f}_i|\bar{e}_i)\delta(\text{start}_i - \text{end}_{i-1} - 1)$$

Approximations:

- distortion limit d: $2^n \to 2^d$
- maximum phrase length m: $n^2 \to n \times m$

- alignment space $O(2^d \times n \times m \times t)$
- weighted derivations $O(2^d \times n \times m \times t \times |\Delta|^{k-1})$
  where $P(E)$ is a k-gram LM components over Δ^*
  and $|\Delta| \propto t \times n$

This space is too large for exact inference

Complete model

$$P(E)P(F,S|E) = \prod_{j=1}^{|E|} \psi(e_j|e_{j-n+1}^{j-1}) \prod_{i=1}^{|S|} \phi(\bar{f}_i|\bar{e}_i)\delta(\mathsf{start}_i - \mathsf{end}_{i-1} - 1)$$

Approximations:

- distortion limit d: $2^n \to 2^d$
- maximum phrase length m: $n^2 \to n \times m$

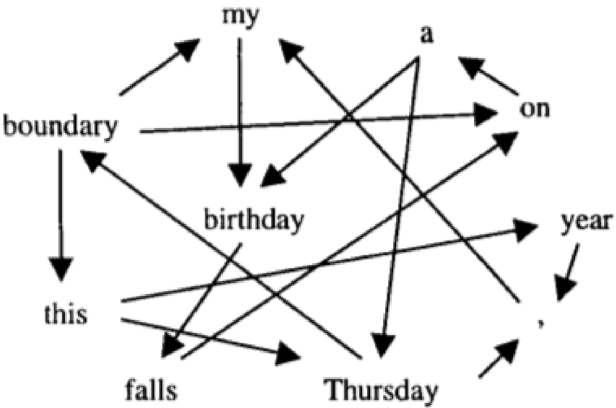
- alignment space $O(2^d \times n \times m \times t)$
- weighted derivations $O(2^d \times n \times m \times t \times |\Delta|^{k-1})$
  where $P(E)$ is a k-gram LM components over Δ^*
  and $|\Delta| \propto t \times n$

This space is too large for exact inference

- pruning: beam search

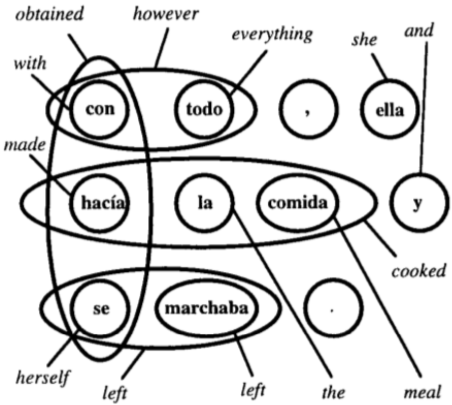
Complexity

[Knight, 1999]



Complexity

[Knight, 1999]



Questions?

References I

John DeNero, Dan Gillick, James Zhang, Dan Klein Why Generative Phrase Models Under perform Surface Heuristics In *Proceedings of the Workshop on Statistical Machine Translation at HLT-NAACL*, pages 31–38, New York, June 2006. Association for Computational Linguistics. URL `http://www.denero.org/content/pubs/naacl06_denero_phrase.pdf`.

John DeNero and Dan Klein. The complexity of phrase alignment problems. In *Proceedings of ACL-08: HLT, Short Papers*, pages 25–28, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Kevin Knight. Decoding complexity in word-replacement translation models. *Comput. Linguist.*, 25(4):607–615, December 1999. URL `http://dl.acm.org/citation.cfm?id=973226.973232`.

References II

Philipp Koehn, Franz Josef Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics. URL http://dx.doi.org/10.3115/1073445.1073462.

Daniel Marcu and Daniel Wong. A phrase-based,joint probability model for statistical machine translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 133–139. Association for Computational Linguistics, July 2002. URL http://www.aclweb.org/anthology/W02-1018.

References III

Markos Mylonakis and Khalil Sima'an. Phrase Translation Probabilities with ITG Priors and Smoothing as Learning Objective. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 133–639, Honolulu, Hawaii, October 2008. URL `http://www.aclweb.org/anthology/D08-1066`

Franz Josef Och and Hermann Ney. Improved Statistical Alignment Models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447, Hong Kong, 2000. URL `http://dx.doi.org/10.3115/1075218.1075274`.

Khalil Sima'an. Computational complexity of probabilistic disambiguation. *Grammars*, 5(2):125–151, 2002. URL `http://dx.doi.org/10.1023/A%3A1016340700671`.

References IV

Mikhail Zaslavskiy, Marc Dymetman, and Nicola Cancedda. Phrase-based statistical machine translation as a traveling salesman problem. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 333–341, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1687878.1687926`.