Phrase-based SMT

Wilker Aziz

Universiteit van Amsterdam
w.aziz@uva.nl

April 9, 2015

Content

1. Statistical model
2. Estimation
3. Inference
4. Conclusions
5. Extensions (further reading)

Noisy channel

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
- $P(F|E)$ ...

Phrase-based model

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

Let's introduce a bidirectional alignment variable

Bidirectional alignment

Example

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

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

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

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

Intuition

$$
\begin{aligned}
P(F|E) &= \sum_A P(A, F|E) \\
&= \sum_A P(\text{segmentation}) \times P(\text{order}) \times P(\text{translation})
\end{aligned}
$$

Example

Intuition

$$P(F|E) = \sum_A P(A, F|E)$$
$$= \sum_A P(\text{segmentation}) \times P(\text{order}) \times P(\text{translation})$$

J'$_1$ ai$_2$ les$_3$ yeux$_4$ noirs$_5$ input

Example

Intuition

$$P(F|E) = \sum_A P(A, F|E)$$
$$= \sum_A P(\text{segmentation}) \times P(\text{order}) \times P(\text{translation})$$

Example	J'$_1$ ai$_2$ les$_3$ yeux$_4$ noirs$_5$	input
	[J'$_1$ ai$_2$] [les$_3$ yeux$_4$] [noirs$_5$]	segmentation

Intuition

$$P(F|E) = \sum_A P(A, F|E)$$
$$= \sum_A P(\text{segmentation}) \times P(\text{order}) \times P(\text{translation})$$

Example	J'$_1$ ai$_2$ les$_3$ yeux$_4$ noirs$_5$	input
	[J'$_1$ ai$_2$] [les$_3$ yeux$_4$] [noirs$_5$]	segmentation
	[J'$_1$ ai$_2$]$_1$ [noirs$_5$]$_2$ [les$_3$ yeux$_4$]$_3$	ordering

Intuition

$$
\begin{aligned}
P(F|E) &= \sum_A P(A, F|E) \\
&= \sum_A P(\text{segmentation}) \times P(\text{order}) \times P(\text{translation})
\end{aligned}
$$

Example	J'_1 ai_2 les_3 yeux_4 noirs_5	input
	$[\text{J'}_1\ \text{ai}_2]$ $[\text{les}_3\ \text{yeux}_4]$ $[\text{noirs}_5]$	segmentation
	$[\text{J'}_1\ \text{ai}_2]_1$ $[\text{noirs}_5]_2$ $[\text{les}_3\ \text{yeux}_4]_3$	ordering
	$[\text{I have}]_1$ $[\text{black}]_2$ $[\text{eyes}]_3$	translation

Model and assumptions

$$
\begin{aligned}
P(F|E) &= \sum_A P(A, F|E) \\
&= \sum_A P(A|E) \times P(F|A, E) \\
&= \sum_A \prod_k \phi(\bar{f}_k|\bar{e}_k)\delta(\text{start}_k - \text{end}_{k-1} - 1)
\end{aligned}
$$

Assumptions

1. uniform alignments (and segmentation)
2. distance-based reordering
3. phrase independence

Estimation

Reordering model

- exponential $\delta(x) = \alpha^{|x|}$

Estimation

Reordering model

- exponential $\delta(x) = \alpha^{|x|}$

Phrase translation model

- EM: requires computing **expected counts** of unseen events (phrase alignments) [Marcu and Wong, 2002]
  DeNero and Klein [2008] proved the problem NP-complete

Estimation

Reordering model

- exponential $\delta(x) = \alpha^{|x|}$

Phrase translation model

- EM: requires computing **expected counts** of unseen events (phrase alignments) [Marcu and Wong, 2002]
  DeNero and Klein [2008] proved the problem NP-complete
- Heuristic: view phrase pairs as **observed** irrespective of context or overlap [Koehn et al., 2003]

Phrase pairs from word alignments

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

Phrase pairs from word alignments

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

- multiple derivations can explain an “observed” phrase pair

Phrase pairs from word alignments

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

- multiple derivations can explain an "observed" phrase pair

Phrase pairs from word alignments

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

- multiple derivations can explain an “observed” phrase pair

Phrase pairs from word alignments

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

- multiple derivations can explain an “observed” phrase pair

Phrase pairs from word alignments

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

- multiple derivations can explain an “observed” phrase pair

Phrase pairs from word alignments

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

- multiple derivations can explain an “observed” phrase pair

Phrase pairs from word alignments

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

- multiple derivations can explain an “observed” phrase pair

Phrase pairs from word alignments

		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

Phrase extraction

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

Phrase extraction

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:

Phrase extraction

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 $\bar{e}$

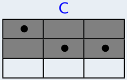
Phrase extraction

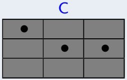
Let $(\bar{f}, \bar{e})$ be a phrase pair

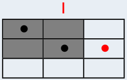
Let A be an alignment matrix

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

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







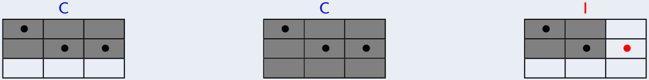
Phrase extraction

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 $\bar{f}$

Phrase extraction

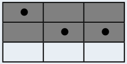
Let $(\bar{f}, \bar{e})$ be a phrase pair

Let A be an alignment matrix

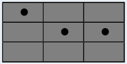
$(\bar{f}, \bar{e})$ consistent with A if, and only if:

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

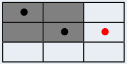
C



C

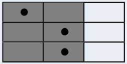


I

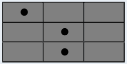


- Words in $\bar{e}$, if aligned, align only with words in $\bar{f}$

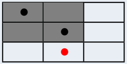
C



C



I



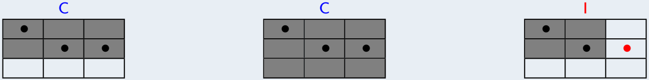
Phrase extraction

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

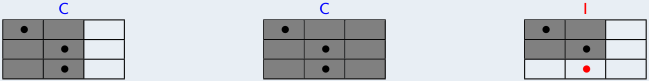
Let A be an alignment matrix

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



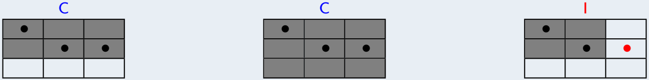
- $(\bar{f}, \bar{e})$ must contain at least one alignment point

Phrase extraction

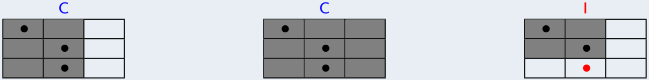
Let $(\bar{f}, \bar{e})$ be a phrase pair
Let A be an alignment matrix

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

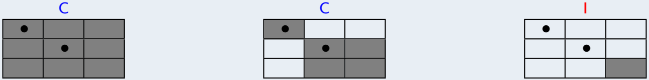
- Words in $\bar{f}$, if aligned, align only with words in $\bar{e}$



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



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



Scoring

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

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

Relative frequency counting

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

Decoding

Disambiguation problem

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

NP-complete [Sima'an, 2002]

Decoding

Disambiguation problem

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

NP-complete [Sima'an, 2002]

Viterbi approximation

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

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]

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]

- distortion limit d: $2^n \rightarrow 2^d$

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]

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

Complete model

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

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

Complete model

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

- alignment space $O(2^d \times n \times t \times m)$
- weighted derivations $O(2^d \times n \times t \times m \times |\Delta|^{k-1})$
  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

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

- alignment space $O(2^d \times n \times t \times m)$
- weighted derivations $O(2^d \times n \times t \times m \times |\Delta|^{k-1})$
  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

Wrap up: model

- relies on Viterbi word-alignments

Wrap up: model

- relies on Viterbi word-alignments
  Is it a good thing?

Wrap up: model

- relies on Viterbi word-alignments
  Is it a good thing?
- underlying model is unclear

Wrap up: model

- relies on Viterbi word-alignments
  Is it a good thing?
- underlying model is unclear
  Do you see why?

Wrap up: model

- relies on Viterbi word-alignments
  Is it a good thing?
- underlying model is unclear
  Do you see why?
- heuristic estimation is straightforward

Wrap up: model

- relies on Viterbi word-alignments
  Is it a good thing?
- underlying model is unclear
  Do you see why?
- heuristic estimation is straightforward
  Can you guess why it works?

Wrap up: complexity

Alignment space

- unconstrained reordering: NP-complete

Wrap up: complexity

Alignment space

- unconstrained reordering: NP-complete
  Do you think it is sensible to allow all permutations?

Wrap up: complexity

Alignment space

- unconstrained reordering: NP-complete
  Do you think it is sensible to allow all permutations?
- tractability requires an ad-hoc distortion limit

Wrap up: complexity

Alignment space

- unconstrained reordering: NP-complete
  Do you think it is sensible to allow all permutations?
- tractability requires an ad-hoc distortion limit
  Do you see the problem here?

Wrap up: reordering

Local

- implicitly modelled within phrases

Wrap up: reordering

Local

- implicitly modelled within phrases
  e.g. probème difficile/difficult problem

Wrap up: reordering

Local

- implicitly modelled within phrases
  e.g. probème difficile/difficult problem
  e.g. ne mange pas/do not eat

Wrap up: reordering

Local

- implicitly modelled within phrases
  e.g. probème difficile/difficult problem
  e.g. ne mange pas/do not eat
  Do you see any problem?

Wrap up: reordering

Local

- implicitly modelled within phrases
  e.g. probème difficile/difficult problem
  e.g. ne mange pas/do not eat
  Do you see any problem?

Nonlocal

- penalised in general

Wrap up: reordering

Local

- implicitly modelled within phrases
  e.g. probème difficile/difficult problem
  e.g. ne mange pas/do not eat
  Do you see any problem?

Nonlocal

- penalised in general
- arbitrarily constrained

Wrap up: reordering

Local

- implicitly modelled within phrases
  e.g. probème difficile/difficult problem
  e.g. ne mange pas/do not eat
  Do you see any problem?

Nonlocal

- penalised in general
- arbitrarily constrained
  Does it suit every language pair?

Discriminative models

Linear model

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

Features

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

Feature weights can be optimised [Och, 2003]

Questions?

References I

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. URL `http://www.aclweb.org/anthology/P/P08/P08-2007`.

Kevin Gimpel and Noah A. Smith. Structured ramp loss minimization for machine translation. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 221–231, Montréal, Canada, June 2012. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/N12-1023`.

References II

Kevin Knight. Decoding complexity in word-replacement translation models. *Comput. Linguist.*, 25(4):607–615, December 1999. ISSN 0891-2017. URL http://dl.acm.org/citation.cfm?id=973226.973232.

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. doi: 10.3115/1073445.1073462. URL http://dx.doi.org/10.3115/1073445.1073462.

References III

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. doi: 10.3115/1118693.1118711. URL http://www.aclweb.org/anthology/W02-1018.

Franz Josef Och. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan, July 2003. Association for Computational Linguistics. doi: 10.3115/1075096.1075117. URL http://www.aclweb.org/anthology/P03-1021.

References IV

Khalil Sima'an. Computational complexity of probabilistic disambiguation. *Grammars*, 5(2):125–151, 2002. ISSN 1386-7393. doi: 10.1023/A:1016340700671. URL http://dx.doi.org/10.1023/A%3A1016340700671.

David A. Smith and Jason Eisner. Minimum risk annealing for training log-linear models. In *Proceedings of the COLING/ACL on Main conference poster sessions*, COLING-ACL '06, pages 787–794, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1273073.1273174.

References V

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. ISBN 978-1-932432-45-9. URL http://dl.acm.org/citation.cfm?id=1687878.1687926.