Statistical Machine Translation: IBM Models 1 and 2

Cognition Volume 116, Issue 2

Michael Collins

[a]Department of Psychology, Rice University, Houston, TX, USA

[b]Department of Chinese, Translation and Linguistics,

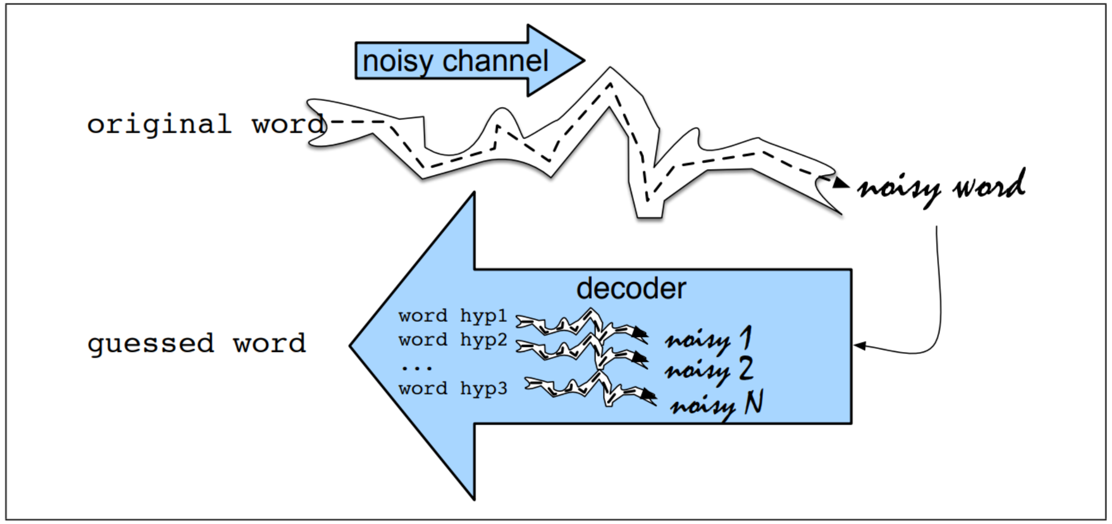
City University of Hong Kong, Kowloon, Hong Kong

Appendix

- **IBM Model**
  - In the noisy-channel approach, the output of the translation model on a new sentence

$$\arg\max_{e\in E} p(e|f) = \arg\max_{e\in E} \frac{p(e)p(f|e)}{\sum_f p(e)p(f|e)} = \arg\max_{e\in E} p(e)p(f|e)$$

  - A language model: a probability $p(e)$ for any sentence $e = e_1 \ldots e_l$ in English
  - A translation model: a conditional probability $p(f|e)$ to any French/English pair of sentences



Noisy Channel Model

- Imagine that the surface form we see is actually a "distorted" form of an original word passed through a noisy channel

- Decoder passes each hypothesis through a model of this channel and picks the word that best matches the surface noisy word

Dan Jurafsky and James H. Martin. Spelling Correction and the Noisy Channel. Speech and Language Processing (3rd ed. draft). 2024. https://web.stanford.edu/~jurafsky/slp3/B.pdf
Michael Collins. Statistical Machine Translation: IBM Models 1 and 2. 2013.

Appendix

- **Alignment**
  - A case where $l = 6, m = 7$
  - Source Sentence $e = \textit{And the programme has been implemented}$
  - Target Sentence $f = \textit{Le programme a ete mis en application}$
  - Alignment variables $a_1, a_2, \ldots, a_7 = \langle 2, 3, 4, 5, 6, 6, 6 \rangle$

Alignment

Le	$\Rightarrow$	the
Programme	$\Rightarrow$	program
a	$\Rightarrow$	has
ete	$\Rightarrow$	been
mis	$\Rightarrow$	implemented
en	$\Rightarrow$	implemented
application	$\Rightarrow$	implemented

Michael Collins. Statistical Machine Translation: IBM Models 1 and 2. 2013.

- **Alignment Models: IBM Model 2**
  - $t(f|e)$: Conditional probability of generating target word f from source word e
  - $q(j|i,l,m)$: Probability of alignment variable a_i (The value j , source & target l & m)

$$p(f_1 \ldots f_m, a_1 \ldots a_m | e_1 \ldots e_l, m) = \prod_{i=1}^{m} q(a_i|i,l,m) t(f_i|e_{a_i})$$

Alignment		
Le	$\Rightarrow$	the
Programme	$\Rightarrow$	program
a	$\Rightarrow$	has
ete	$\Rightarrow$	been
mis	$\Rightarrow$	implemented
en	$\Rightarrow$	implemented
application	$\Rightarrow$	implemented

$$\begin{aligned} & p(f_1 \ldots f_m, a_1 \ldots a_m | e_1 \ldots e_l, m) \\ = \; & q(2|1,6,7) \times t(\mathit{Le}|\mathit{the}) \\ & \times q(3|2,6,7) \times t(\mathit{Programme}|\mathit{program}) \\ & \times q(4|3,6,7) \times t(\mathit{a}|\mathit{has}) \\ & \times q(5|4,6,7) \times t(\mathit{ete}|\mathit{been}) \\ & \times q(6|5,6,7) \times t(\mathit{mis}|\mathit{implemented}) \\ & \times q(6|6,6,7) \times t(\mathit{en}|\mathit{implemented}) \\ & \times q(6|7,6,7) \times t(\mathit{application}|\mathit{implemented}) \end{aligned}$$

Michael Collins. Statistical Machine Translation: IBM Models 1 and 2. 2013.

Appendix

- **Independence Assumptions in IBM Model 2**
  - The chain rule of probabilities to decompose this into two terms

$$P(F_1 = f_1 \ldots F_m = f_m, A_1 = a_1 \ldots A_m = a_m | E_1 = e_1 \ldots E_l = e_l, L = l, M = m)$$

$$= P(A_1 = a_1 \ldots A_m = a_m | E_1 = e_1 \ldots E_l = e_l, L = l, M = m)$$
$$\times P(F_1 = f_1 \ldots F_m = f_m | A_1 = a_1 \ldots A_m = a_m, E_1 = e_1 \ldots E_l = e_l, L = l, M = m)$$

  - Independence assumptions $q(a_i | i, l, m)$
    - It is independent of the English words $E_1 \ldots E_l$, and of the other alignment variables

$$P(A_1 = a_1 \ldots A_m = a_m | E_1 = e_1 \ldots E_l = e_l, L = l, M = m)$$
$$= \prod_{i=1}^{m} P(A_i = a_i | A_1 = a_1 \ldots A_{i-1} = a_{i-1}, E_1 = e_1 \ldots E_l = e_l, L = l, M = m)$$
$$= \prod_{i=1}^{m} P(A_i = a_i | L = l, M = m)$$

$$P(A_i = a_i | L = l, M = m) = q(a_i | i, l, m)$$

Michael Collins. Statistical Machine Translation: IBM Models 1 and 2. 2013.

Appendix

- **Independence Assumptions in IBM Model 2**
  - The chain rule of probabilities to decompose this into two terms

$$P(F_1 = f_1 \dots F_m = f_m, A_1 = a_1 \dots A_m = a_m | E_1 = e_1 \dots E_l = e_l, L = l, M = m)$$

$$= P(A_1 = a_1 \dots A_m = a_m | E_1 = e_1 \dots E_l = e_l, L = l, M = m)$$
$$\times P(F_1 = f_1 \dots F_m = f_m | A_1 = a_1 \dots A_m = a_m, E_1 = e_1 \dots E_l = e_l, L = l, M = m)$$

  - Independence assumptions $t(f_i|e_{a_i})$
    - It is independent of the English words $E_1 \dots E_l$, and of the other alignment variables

$$P(F_1 = f_1 \dots F_m = f_m | A_1 = a_1 \dots A_m = a_m, E_1 = e_1 \dots E_l = e_l, L = l, M = m)$$
$$= \prod_{i=1}^{m} P(F_i = f_i | F_1 = f_1 \dots F_{i-1} = f_{i-1}, A_1 = a_1 \dots A_m = a_m, E_1 = e_1 \dots E_l = e_l, L = l, M = m)$$
$$= \prod_{i=1}^{m} P(F_i = f_i | E_{a_i} = e_{a_i})$$

$$P(F_i = f_i | E_{a_i} = e_{a_i}) = t(f_i | e_{a_i})$$

Michael Collins. Statistical Machine Translation: IBM Models 1 and 2. 2013.

Appendix

- **Alignment**
  - A case where $l = 6, m = 7$
  - Source Sentence $e = \textit{And the programme has been implemented}$
  - Target Sentence $f = \textit{Le programme a ete mis en application}$
  - Alignment variables $a_1, a_2, \ldots, a_7 = \langle 2, 3, 4, 5, 6, 6, 6 \rangle$

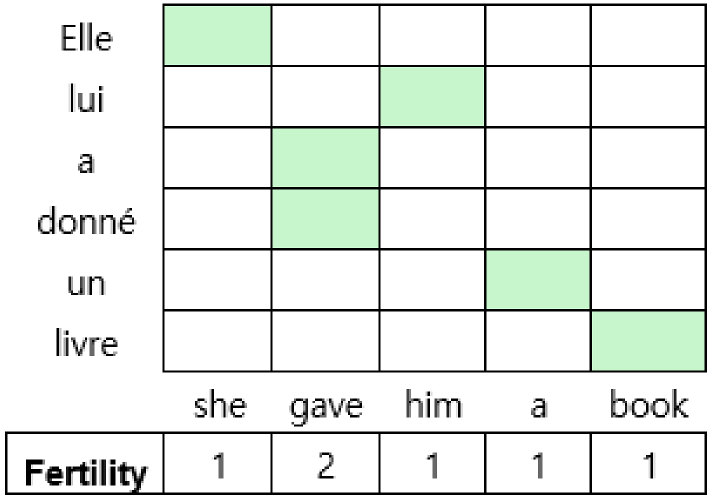
Alignment		
Le	$\Rightarrow$	the
Programme	$\Rightarrow$	program
a	$\Rightarrow$	has
ete	$\Rightarrow$	been
mis	$\Rightarrow$	implemented
en	$\Rightarrow$	implemented
application	$\Rightarrow$	implemented

Michael Collins. Statistical Machine Translation: IBM Models 1 and 2. 2013.

Appendix

- **Alignment Models: IBM Model 3**
  - Mapping relationship between source and target sentence
  - Fertility
    - The number of target words generated by a single source word
  - Distortion
    - How the position of words changes from source to target

Phrase alignment (many-to-many alighment)



- Fertility is modeled using a probability distribution $n(\phi|f)$
- ϕ = the number of target words produced by the source word f

- Distortion is represented by a probability distribution $d(j|i,l,m)$
- A word at position i in the source sentence
- A word at position j in the target sentence
- The lengths l and m of the source and target sentences

Peter Brown et al. The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics, 19(2):263–311, 1993.