Machine Translation: Summer Term 2021

Ex2: SMT Intuition; Probability and Noisy Channel.

Sangeet Sagar (7009050)
sasa00001@stud.uni-saarland.de

June 17, 2021

1. **What is great about RBMT?**
   RBMT is capable to achieve very high accuracy on a closed dataset. It does not need bilingual texts or parallel dataset. Another benefit of RBMT is that it is domain independent. Its rules are framed in such a way that it is valid for a sentence from any domain (certain exception might subject the need to write new rules). We also have a total control on the translations generated becuase all rules are hand made, hence easy to debug.

2. **What is not so good about RBMT?**
   RBMT is notoriously famous for the need of large number of rules that leads to a complex system. It can only capture local phenomena and any long range contexts are hard to capture. It is also hard to deal with rule interactions in big systems, ambiguity, and idiomatic expressions [Source: Wikipedia].

3. **Why do we try to use machine learning or statistical estimation from data for MT?**

4. **In your own words, explain**

$$\hat{e} = \underset{e}{\operatorname{argmax}}\, P(e|f)$$

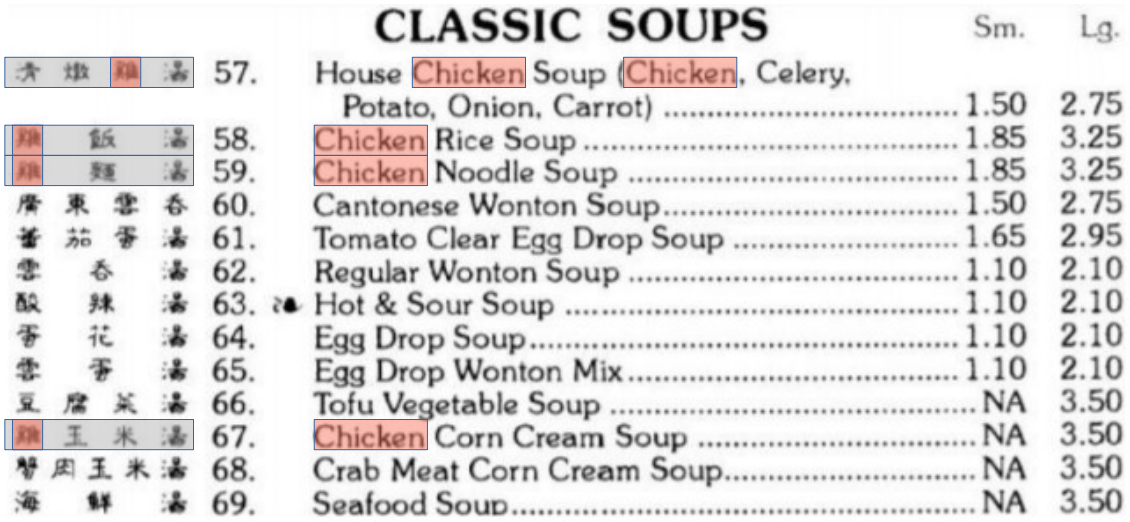
   This is the fundamental equation of Statistical Machine Translation (SMT) in terms of probability. The goal is to model a probability distribution P that finds the most probable English sentence e given a foreign language sentence f i.e. it should tell the probability of all possible English sentences that could be a translation of the given foreign language sentence. In the space of all possible English sentence, $argmax$ searches for a sentence that maximizes this probability and the resulting (English) sentence is the best translation $\hat{e}$ of the input sentence.

5. **What kinds of data do we need for SMT?**

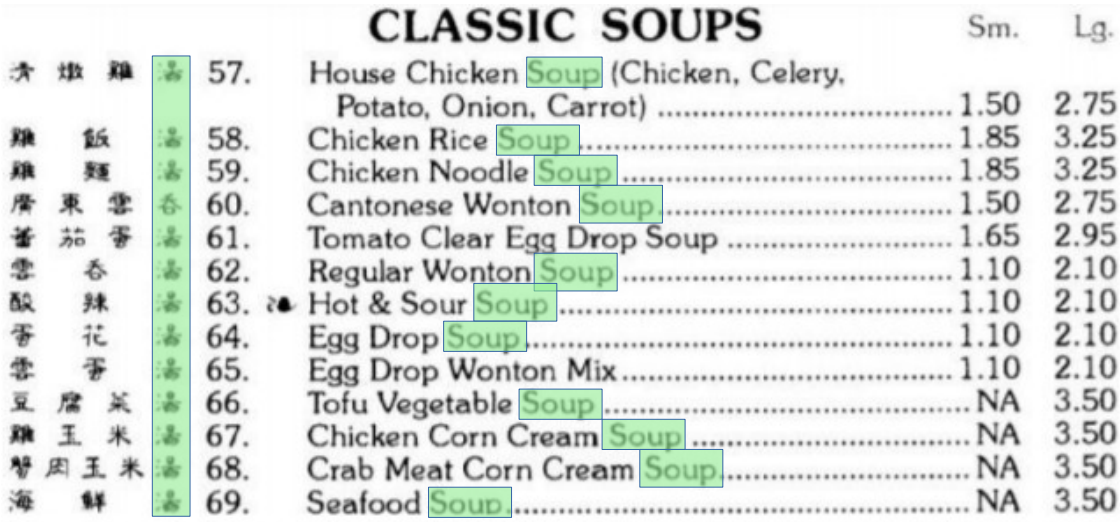
   - Human translation: Bi-text, parallel corpus of the language pair.
   - Monolingual data of the target language.

6. **Given this bitext, which symbol is the likely Chinese symbol for *chicken*? Which symbol is the likely symbol for *soup*?**

		CLASSIC SOUPS	Sm.	Lg.
雞	57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
雞 飯 湯	58.	Chicken Rice Soup	1.85	3.25
雞 麵 湯	59.	Chicken Noodle Soup	1.85	3.25
廣東雲吞	60.	Cantonese Wonton Soup	1.50	2.75
番茄蛋花湯	61.	Tomato Clear Egg Drop Soup	1.65	2.95
雲吞 湯	62.	Regular Wonton Soup	1.10	2.10
酸辣 湯	63.	Hot & Sour Soup	1.10	2.10
蛋花 湯	64.	Egg Drop Soup	1.10	2.10
雲吞 湯	65.	Egg Drop Wonton Mix	1.10	2.10
豆腐菜湯	66.	Tofu Vegetable Soup	NA	3.50
雞玉米湯	67.	Chicken Corn Cream Soup	NA	3.50
蟹肉玉米湯	68.	Crab Meat Corn Cream Soup	NA	3.50
海鮮 湯	69.	Seafood Soup	NA	3.50



		Sm.	Lg.
	CLASSIC SOUPS		
57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
58.	Chicken Rice Soup	1.85	3.25
59.	Chicken Noodle Soup	1.85	3.25
60.	Cantonese Wonton Soup	1.50	2.75
61.	Tomato Clear Egg Drop Soup	1.65	2.95
62.	Regular Wonton Soup	1.10	2.10
63.	Hot & Sour Soup	1.10	2.10
64.	Egg Drop Soup	1.10	2.10
65.	Egg Drop Wonton Mix	1.10	2.10
66.	Tofu Vegetable Soup	NA	3.50
67.	Chicken Corn Cream Soup	NA	3.50
68.	Crab Meat Corn Cream Soup	NA	3.50
69.	Seafood Soup	NA	3.50



		Sm.	Lg.
	CLASSIC SOUPS		
57.	House Chicken Soup (Chicken, Celery, Potato, Onion, Carrot)	1.50	2.75
58.	Chicken Rice Soup	1.85	3.25
59.	Chicken Noodle Soup	1.85	3.25
60.	Cantonese Wonton Soup	1.50	2.75
61.	Tomato Clear Egg Drop Soup	1.65	2.95
62.	Regular Wonton Soup	1.10	2.10
63.	Hot & Sour Soup	1.10	2.10
64.	Egg Drop Soup	1.10	2.10
65.	Egg Drop Wonton Mix	1.10	2.10
66.	Tofu Vegetable Soup	NA	3.50
67.	Chicken Corn Cream Soup	NA	3.50
68.	Crab Meat Corn Cream Soup	NA	3.50
69.	Seafood Soup	NA	3.50

7. **Given the following bitext and word alignment (indicated in terms of colour codes), (i) estimate a word based probabilistic translation dictionary (a translation model), (ii) find**

the best "translations" of

- They love the girl
- I talk to the dog

into French under the model and (iii) compute the probabilities for the best translations under the model based on the word-based translation probabilities, assuming that the probabilities are independent of each other:

$$P(f_1 f_2 \dots f_n) = \prod_{i=1}^{n} P(f_i|e_i)$$

Given the above aligned data, we prepare the collated statistics-

English word	Foreign word	Frequency	Probablity
I	J'	2	0.67
	Je	1	0.33
love	aime	2	0.67
	aiment	1	0.33
the	le	3	0.60
	la	2	0.40
boy	garçon	1	1
girl	fille	1	1
mother	mère	1	1
dog	chiene	3	1
they	lls	3	1
talk	parlent	2	0.67
	parle	1	0.33
to	à	2	0.67
	au/_the	1	0.33

1)

They	love	the	girl
lls	aime	le	fille
1	0.67	0.60	1
lls	aiment	la	fille
1	0.33	0.40	1

But from the parallel data, we see that lls occurs with aiment. Hence the choice aime is ruled out by the language model. Therefore the appropriate translation would be:

They	love	the	girl
lls	aiment	le	fille
1	0.33	0.60	1

P(lls aiment le fille) = 1*0.33*0.60*1 = 0.198

2)

I	talk	to	the	dog
J'	parlent	à	le	chiene
0.67	0.67	0.67	0.60	1

Je	parle	au	la	chiene
0.33	0.33	0.33	0.40	1

From the parallel data, we see that Je occurs with parle. Hence the choice parlent is ruled out by the language model. So, Je parle is a good translation. Now, parle occurs together with à twice and with au just once. Hence, parle à is a more suitable choice than parle au. Also, à la is seen to occur together in the aligned data, hence the other possibility i.e. à le is also ruled out. Therefore, the most appropriate translation would be-

I	talk	to	the	dog
Je	parle	à	la	chiene
0.33	0.33	0.67	0.40	1

P(Je parle à la chiene) = 0.33*0.33*0.67*0.40*1 = 0.0291

8. **Given a sequence of n numbers/measurements/numerical observations $x_1, x_2, \ldots x_n$, please define**

   - **Population mean, sample mean**
   - **Population variance, sample variance**
   - **Population standard deviation, sample standard deviation**

   - Population mean is computed over the entire population size n while sample mean is computed over $n-1$ where n is the sample size.
     Population mean
     $$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$
     Sample mean
     $$\bar{x} = \frac{1}{n-1}\sum_{i=1}^{n} x_i$$
   - Population variance
     $$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$$
     Sample variance
     $$Var(x) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$
   - Population deviation
     $$\sigma = \sqrt{\sigma^2} = \sqrt{Var(x)}$$
     Sample deviation
     $$s = \sqrt{s^2} = \sqrt{Var(x)}$$

9. **Explain the notions of sample space, outcome and event in set-based formalisations of probability.**
Sample space (Ω) is the set of all possible outcomes of an event. It defines

$$\Omega = \{\omega_1, \omega_2, \cdots, \omega_n\}$$

An outcome is a possible result of an experiment or trial represented by $\omega_1, \omega_2, \cdots, \omega_n$. An event E is composed by basic outcomes ω_i and is a subset of sample space. $E \subseteq \Omega$

10. **What is a Laplace experiment and a Laplace probability?**
Laplace experiments have equiprobable events due to its property of symmetricity where each outcome is equally as others. Laplace probablity is given as

$$P(A) = \frac{\#\text{outcomes of event A}}{\#\text{total outcomes}}$$

11. **Probabilities can be estimated from counts (relative frequencies). Give the following observations, estimate the probabilities:**

	Car	**Lorry**	**M-cycle**	**Cycle**	**Pedes.**	**n**
$\#(x)$	124	49	7	64	271	515
$\frac{\#x}{n}$	$\frac{124}{515}$	$\frac{49}{515}$	$\frac{7}{515}$	$\frac{64}{515}$	$\frac{271}{515}$	-
$P(x)$	0.24	0.095	0.013	0.124	0.526	-

12. **Please complete the following table describing the Boolean Algebra of events:**

- $A \cap B = B \cap A$
- $A \cap (B \cap C) = (A \cap B) \cap (A \cap C)$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- $A \cap A = A$
- $A \cap (A \cup B) = (A \cap A) \cup (A \cap B) = A$
- $\overline{A \cap B} = \bar{A} \cup \bar{B}$
- $A \cap \phi = \phi$
- $A \cap \bar{A} = \phi$
- $\bar{\bar{A}} = A$

- $A \cup B = B \cup A$
- $A \cup (B \cup C) = (A \cup B) \cup (A \cup C)$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
- $A \cup A = A$
- $A \cup (A \cap B) = (A \cup A) \cap (A \cup B) = A$
- $\overline{A \cup B} = \bar{A} \cap \bar{B}$
- $A \cup \Omega = \Omega$
- $A \cup \bar{A} = \phi$

13. **Define conditional probability:**

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)}$$

14. **Given a fair 6-sided dice, what are**

- $P(odd) = P(1, 3, 5) = \frac{3}{6}$
- $P(odd|prime) = \frac{P(odd \cap prime)}{P(prime)} = \frac{2/6}{3/6} = \frac{2}{3}$
- $P(prime|odd) = \frac{P(prime \cap odd)}{P(odd)} = \frac{2/6}{3/6} = \frac{2}{3}$
- $P(even|prime) = \frac{P(even \cap prime)}{P(prime)} = \frac{1/6}{3/6} = \frac{1}{3}$
- $P(\{5\}|odd) = \frac{P(\{5\} \cap odd)}{P(odd)} = \frac{1/6}{3/6} = \frac{1}{3}$
- $P(\{2, 5\}|odd) = \frac{P(\{2,5\} \cap odd)}{P(odd)} = \frac{1/6}{3/6} = \frac{1}{3}$

- $P(\phi|odd) = \frac{P(\phi \cap odd)}{P(odd)} = \frac{0}{3/6} = 0$

15. **When are two events mutually exclusive, when are two events independent?**
Two events are mutually exclusive if they can not occur at the same time. Two events are independent if occurrence of one does not affect the probability of occurrence of the other.

16. **Give the specific and the general version of the addition rule of probabilities:**

- $P(A \cup B) = P(A) + P(B)$ (mutually exclusive events)
- $P(A \cup B) = P(A) + P(B) - P(A \cup B)$

17. **Give the specific and the general version of the multiplication rule of probabilities:**

- $P(A \cap B) = P(B) \cdot P(A)$
- $P(A \cap B) = P(B) \cdot P(A|B) = P(A) \cdot P(B|A)$

18. **Give the complement rule of probability:**
$P(\bar{a}) = 1 - P(A)$

19. **Expand the following using the chain rule of probability**

$$\begin{aligned} P(w_1, w_2, \cdot, w_n) &= P(w_1)P(w_2|w_1)P(w_3|w_2, w_1)\cdot \\ &= \prod_{i=1}^{n} P(w_i|w_{i-1}, w_{i-2}, \cdot, w_1) \end{aligned}$$

20. **What is the prior, the likelihood and the posterior in Bayes rule:**

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Where:

- $P(A|B)$: Posterior
- $P(B|A)$: Likelihood
- $P(A)$: Prior

21. **Prove Bayes Rule**
Using properties of conditional probability

$$P(A|B) = \frac{P(A,B)}{P(B)} P(A,B) = P(A|B) \cdot P(B) \tag{1}$$

Similarly

$$P(B,A) = P(B|A) \cdot P(A) \tag{2}$$

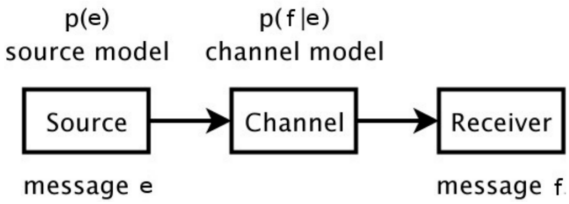
Equating above two as $P(A,B) = P(B,A)$, we get

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A) P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{3}$$

22. **Why is Bayes rule useful?**
Bayes theorem provides a way to revise existing predictions or theories (update probabilities) given new or additional evidence [Hayes].

23. **In your own words, relate the fundamental rule of statistical machine translation (SMT) to the noisy channel model (NC):**
What is the translation model, what is the language model, what is the source model, what is the channel model, what is the prior, the likelihood and the posterior? In what sense is this a MAP (maximum a posteriori) decision rule?

- $P(e)$: the source mode, the language model
- $P(f|e)$: the channel model, the translation model

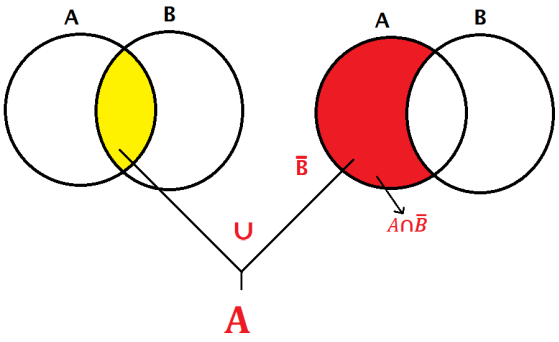


Given source (encoder), channel and receiver (decoder), the goal is to recover the original message. We observe some distorted message in french (f) given an input message in English (e) and the information about this distortion is stored in the translation model $P(f|e)$.

The original message is recovered as

$$\begin{aligned}\hat{e} &= \underset{e}{\operatorname{argmax}}\, p(e|f) \\ &= \underset{e}{\operatorname{argmax}}\, \frac{p(f|e)\cdot p(e)}{p(f)} \\ &= \underset{e}{\operatorname{argmax}}\, p(f|e)\cdot p(e) \quad (p(f) \text{ is constant and does not help in maximizing in this probability})\end{aligned}$$

24. **Rule of total probability: can you show in terms of a drawing why** $P(A) = P(A \cap B) \cup P(A \cap \bar{B})$



25. **Total probability: why is** $P(A \cap B) \cup P(A \cap \bar{B}) = P(A|B)\cdot P(B) + P(A|\bar{B})\cdot P(\bar{B})$

$$\begin{aligned}P(A \cap B) \cup P(A \cap \bar{B}) &= P(A|B)\cdot P(B) \cup P(A|\bar{B})\cdot P(\bar{B}) \quad \text{(multiplication rule)} \\ &= P(A|B)\cdot P(B) + P(A|\bar{B})\cdot P(\bar{B}) \quad \text{(specific version of addition rule)}\end{aligned}$$

References

Adam Hayes. Bayes' theorem definition. `https://www.investopedia.com/terms/b/bayes-theorem.asp`. Updated: 2020-06-14.