STATISTICAL MACHINE TRANSLATION

LING83600: Language Technology

OUTLINE

- Some basic concepts in machine translation design
- Evaluating translation quality using BLEU score
- The generative models underlying Candide, the influential statistical machine translation system

THE NOISY CHANNEL MODEL OF TRANSLATION

Warren Weaver, 1949 Rockefeller Foundation memorandum *Translation*:

“When I look at an article in Russian, I say: this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode.”

$\mathrm{argmax}_e \, P(e) \, P(r \mid e)$

Machine translation received massive US government funding in the ‘50s and early ‘60s, but made next to no progress on the core problems.

The ALPAC report (1964) recommended that government-funded MT research focus on:

1. <u>practical methods for evaluation of translations</u>

…

3. <u>evaluation of quality and cost of various sources of translations</u>

…

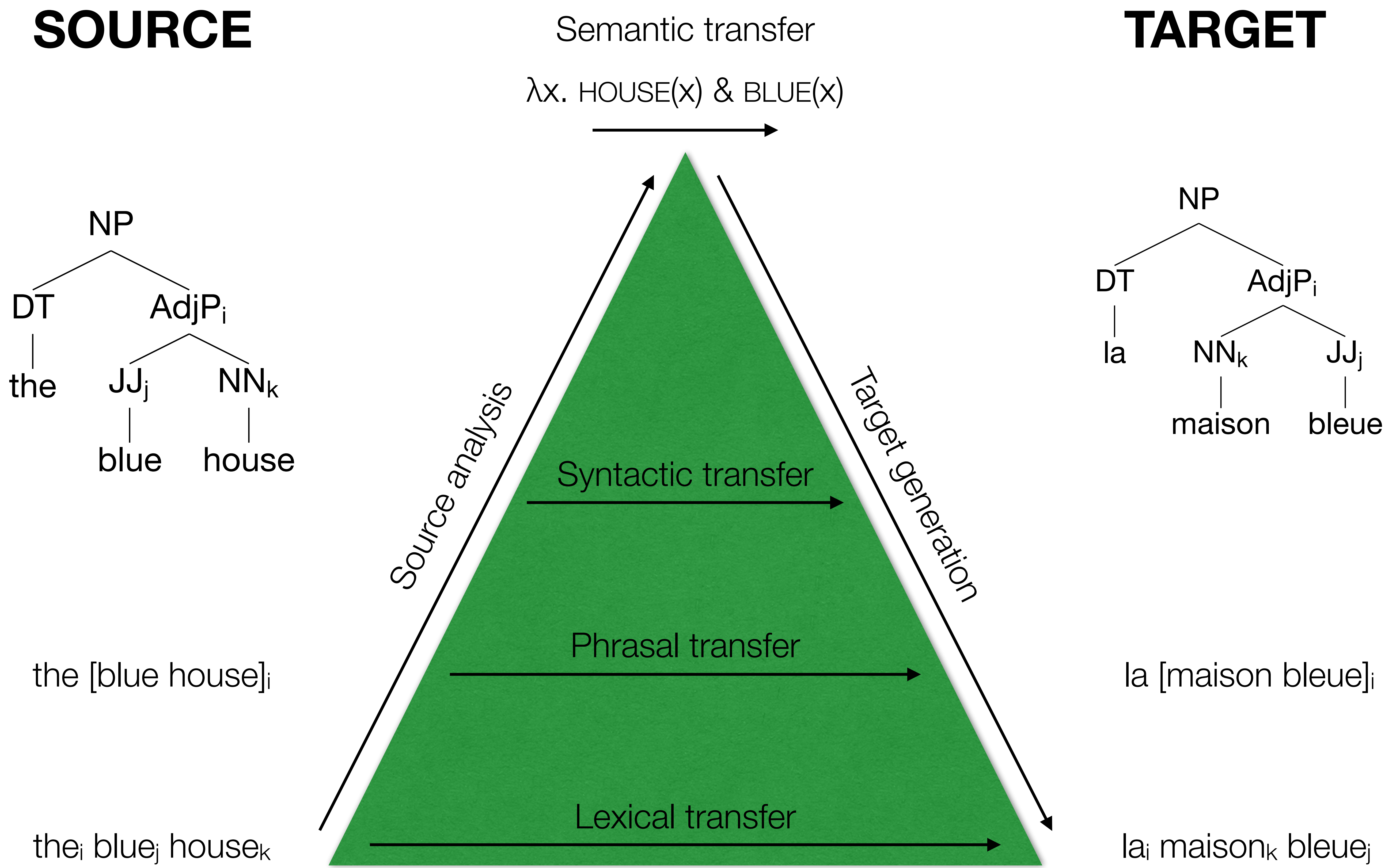
9. <u>production of adequate reference works for the translator, including the adaptation of glossaries that now exist primarily for automatic dictionary look-up in machine translation</u>

This (ultimately) lead researchers to adopt a clearer problem statement, the modeling of *translator behavior*.

Effective domain-general machine translation systems consist of…

data-driven, language-agnostic models of *translator behavior…*

paired with language-specific models of *linguistic analysis* and *generation*.



VAUQUOIS TRIANGLE

THE QUADRATIC GROWTH PROBLEM

As the number of languages a system supports (*n*) increases, the number of translation models needed grows quadratically to $n^2 - n$*

Thus, when developing multilingual translation systems, we place language-specific methods in the monolingual *analysis* and *generation* models so the *translation* model is as language-independent as possible.

*Note that translation models need not be invertible.

In the early 1990s, a team at IBM Research built **Candide**, the first modern *statistical* machine translation system. We will be reviewing the intuitions behind Candide in great detail.

BLEU SCORE HANDOUT

Candidate:
Many will lose their right to a pension in their own name because of their husband ’s income .

Reference: *Many will lose their right to draw a pension with their own name because of the income of their husband .*

Candidate:
Many will lose their right to a pension in their own name because of their husband ’s income .

Reference: *Many will lose their right to draw a pension with their own name because of the income of their husband .*

Candidate:
Many will lose their right to a pension in their own name because of their husband 's income .

Reference: *Many will lose their right to draw a pension with their own name because of the income of their husband .*

p_1: 17 / 19 = .895
p_2: 12 / 18 = .667
p_3: 8 / 17 = .471
p_4: 5 / 16 = .313

GM_n = .544
BP = .900
$BLEU$ = .490

BLEU is one of the first evaluation metrics which is well-correlated with human judgements of translation quality.

THE CANDIDE STATISTICAL MACHINE TRANSLATION MODELS

[Brown et al. 1990, 1993, Knight 1999]

TRANSLATION STORY ELEMENTS

- The translation model $P(t \mid s)$ helps to select likely translations:

  $P(house \mid maison) > P(dog \mid maison)$

- The language model $P(t)$ helps with source-to-target polysemy:

  $P(\textit{in the end zone}) > P(\textit{on the end zone})$

- It also helps to sort out word order:

  $P(\textit{the dog runs}) > P(\textit{runs dog the})$

- Decoding helps us find “likely” “stories”.

MODEL I: BASIC STORY

1. Given a source S of length $|S|$, select a target length $|T|$ according to $P(|T| \mid |S|)$

2. Populate T with tokens t according to $P(t \mid s)$

3. Reorder the tokens in T to maximize $P(t_0 \ldots t_{|T|})$

MODEL I: TRANSLATION MODEL ESTIMATION VIA THE EM ALGORITHM

1. Compute $P(t|\, s)$, the MLE conditional probability distribution of s and t co-occurring

2. For n iterations:

    1. Initialize $a(s,\ t) = 0$, $Z(t) = 0$ for all $s \in S$, $t \in T$.

    2. For all pairs of sentences S, T:

       For all $s \in S$, $t \in T$,

       $$a(s, t) = a(s, t) + P(t \mid s)$$
       $$Z(t) = Z(t) + P(t \mid s)\ .$$

    3. For all $s,\ t$, let

       $P(t \mid s) = a(s,\ t) \ /\ Z(t)$

       then normalize $P(t \mid s)$.

Source:	**Target:**
LA MAISON BLEUE	*THE BLUE HOUSE*
LA MAISON	*THE HOUSE*
MAISON	*HOUSE*

ITERATION 0 (MLE ONLY)

$P(\text{HOUSE} \mid \text{MAISON}) = .500$

$P(\text{BLUE} \mid \text{MAISON}) = .167$

$P(\text{THE} \mid \text{MAISON}) = .333$

ITERATION 1

$P(\text{HOUSE} \mid \text{MAISON}) = .440$

$P(\text{BLUE} \mid \text{MAISON}) = .233$

$P(\text{THE} \mid \text{MAISON}) = .327$

ITERATION 2

$P(\text{HOUSE} \mid \text{MAISON}) = .478$

$P(\text{BLUE} \mid \text{MAISON}) = .196$

$P(\text{THE} \mid \text{MAISON}) = .325$

ITERATION 10

$P(\text{HOUSE} \mid \text{MAISON}) = .643$

$P(\text{BLUE} \mid \text{MAISON}) = .077$

$P(\text{THE} \mid \text{MAISON}) = .280$

MODEL II

1. Given a source S of length $|S|$, select a target length $|T|$ according to $P(|T| \mid |S|)$

2. For each source token s_i and the null token, "align" it with some t_j according to $P(i, j)$

3. Translate all aligned source/target s_i, t_j pairs according to $P(t_j \mid s_i)$.

MODEL III

Distortion parameters are now sensitive to lengths:

$P(i \mid j, |S|, |T|)$ is the probability that source token j corresponds with (i.e., is aligned to and is translated by) target token i when the source is $|S|$ tokens long and the target is $|T|$ tokens long

Each source word has a *fertility* parameter:

$P(3 \mid s)$ is the probability that s aligns to exactly 3 target words

MODEL III

$P(n \mid s)$: target token s aligns to n source tokens

$P(t \mid \varnothing)$: a target token t aligns to no source token

$P(t \mid s)$: target token t is generated by aligned source token s

$P(j \mid i, |S|, |T|)$: target token t appears in position j when it is generated by aligned source token in position i and the source and target are $|S|$ and $|T|$ long, respectively

$P(t_0 \ldots t_{|T|})$: the target consists of $t_0 \ldots t|_T|$

MODEL III

Mary	did	not		slap			the	green	witch
Mary		not	slap	slap	slap		the	green	witch
Mary		not	slap	slap	slap	∅	the	green	witch
Maria		no	dió	una	bofetada	a	la	verde	bruja
Maria		no	dió	una	bofetada	a	la	bruja	verde

[h/t: Kevin Knight]

EM ALGORITHM FOR MODEL III

- One candidate alignment:

  $f_d(8 \mid 5, 7, 9) = f_d(8 \mid 5, 7, 9) + 1\ldots$

- Two candidate alignments:

  $f_d(8 \mid 5, 7, 9) = f_d(8 \mid 5, 7, 9) + 1/2$
  $f_d(8 \mid 6, 7, 9) = f_d(8 \mid 6, 7, 9) + 1/2$

- But, the set of possible alignments grows *very fast* so we use *Viterbi training* rather than all possible alignments.

BASIC PHRASE-BASED TRANSLATION MODELS

1. Segment source S into phrases $s_1 \ldots s_N$
2. Reorder each s_i according to distortion model
3. Translate each s_i according to phrasal translation model

[Och & Ney 2004]

PHRASAL ALIGNMENTS

Maria		no	[dió	una	bofetada]	a	la	bruja	verde
Mary	did	not		slap			the	green	witch

PHRASAL ALIGNMENT TEMPLATES WITH GAPS

… ne voudrais pas voyager par chemin de fer

… would not like traveling by railroad

[Bansal et al. 2011]

HIERARCHICAL PHRASAL ALIGNMENT

he adores [listening [to music]$_j$]$_i$

kare ha [[ongaku wo]$_j$ kiku]$_i$ no ga daisuki desu

[Knight & Yamada 2001, Chiang 2005]

OPEN-SOURCE SOFTWARE

- EGYPT (CSLP/JHU 1999 team): IBM models I-V
- GIZA++ (Och & Ney 2003): optimized IBM models
- MOSES (Koehn 2009): IBM model "VI" onward...

FURTHER READING

Kevin Knight. 1999. *A statistical MT workbook.* Ms., University of Southern California.

Philipp Koehn. 2010. *Statistical Machine Translation*. Oxford University Press.

Peter Brown, Vincent della Pietra, Stephen della Pietra, and Robert Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics* 19(2): 263-312.