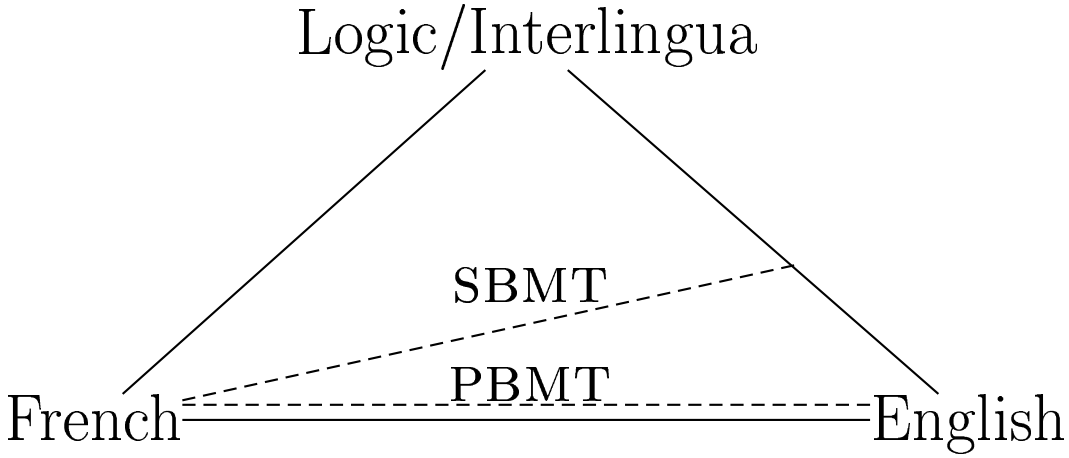
Statistical Machine Translation

Gerald Penn

CSC 401
University of Toronto

http://www.cs.toronto.edu/~gpenn/csc401

The Vauquois Triangle (1968)



A very influential geometric depiction of a quandry that haunted machine translation research for years: at what level of linguistic structure should the "transfer" be made?

Current statistical machine translation research is roughly divided into two camps: *phrase-based* and *syntax-based*. We'll cover phrase-based MT.

The IBM Translation Model

We want to find the best English translation of a given French sentence:

$$\hat{E} = \underset{E}{\text{argmax}}\ p(E|F)$$

Apply Bayes's Rule:

$$\hat{E} = \underset{E}{\text{argmax}}\ p(F|E) \cdot p(F)$$

. . . and for efficiency, make the assumption that there is a single very probable *alignment*, A, between the two strings:

$$\hat{E} \approx \underset{E}{\text{argmax}}\ p(F, A|E) \cdot p(F)$$

(without this assumption, how would $p(F|E)$ relate to $p(F, A|E)$?)

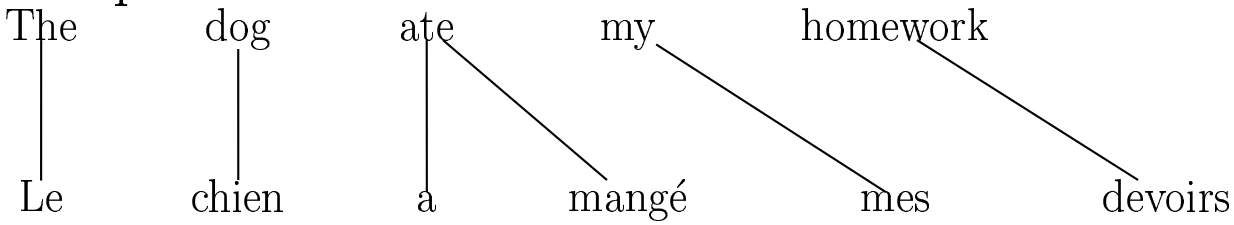
Just as in ASR, $p(F)$ is a *language model*. $p(F, A|E)$ is called a *translation model*.

Alignments

An alignment is a mapping from the substrings of one string to the substrings of another. There are a lot of these.

So we restrict them, usually by limiting the sizes of the subsets that can correspond, often by limiting the *skew* of the correspondence between them.

Examples:



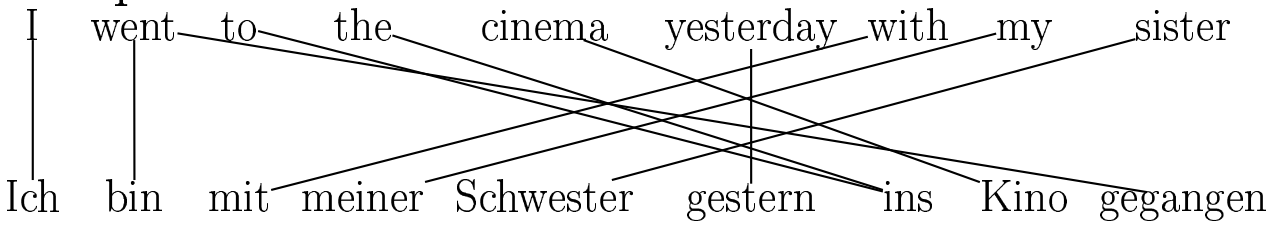
Here, for each French word, there is a unique English word that it is aligned with (size of substring on the French side is always 1), but not *vice versa*. Very little skew.

Alignments

An alignment is a mapping from the substrings of one string to the substrings of another. There are a lot of these.

So we restrict them, usually by limiting the sizes of the subsets that can correspond, often by limiting the *skew* of the correspondence between them.

Examples:



Same here for German, except lots of skew. The word order is almost completely reversed.

Calculating Aligned Translation Model Probability

$$p(F, A|E) = \prod_{i=1}^{|E|} p(n(e_i)|e_i) \cdot \prod_{j=1}^{|F|} p(f_j|e_{a_j}) \cdot d(A|E, F)$$

where:

- $n(e_i)$ is the *fertility* of e_i. It indicates how many French words are aligned with e_i.
- $p(f_j|e_{a_j})$ is a *lexical transfer probability*. It indicates which words in English are likely to be aligned with which words in French.
- $d(A|F, E)$ is a *distortion probability*. It indicates the probability of a particular order of French words, given the words and their English counterparts.

We must optimize this function subject to the constraint that $\sum_{i=1}^{|E|} n(e_i) = |F|$.

Calculating Aligned Translation Model Probability

$$p(F, A|E) = \prod_{i=1}^{|E|} p(n(e_i)|e_i) \cdot \prod_{j=1}^{|F|} p(f_j|e_{a_j}) \cdot d(A|E, F)$$

where:

- $n(e_i)$ is the *fertility* of e_i. It indicates how many French words are aligned with e_i.
- $p(f_j|e_{a_j})$ is a *lexical transfer probability*. It indicates which words in English are likely to be aligned with which words in French.
- $d(A|F, E)$ is a *distortion probability*. It indicates the probability of a particular order of French words, given the words and their English counterparts.

Fertility and lexical transfer are estimated by training on bilingual corpora that are annotated with alignments (*bitexts*).

Calculating Aligned Translation Model Probability

$$p(F, A|E) = \prod_{i=1}^{|E|} p(n(e_i)|e_i) \cdot \prod_{j=1}^{|F|} p(f_j|e_{a_j}) \cdot d(A|E, F)$$

where:

- $n(e_i)$ is the *fertility* of e_i. It indicates how many French words are aligned with e_i.
- $p(f_j|e_{a_j})$ is a *lexical transfer probability*. It indicates which words in English are likely to be aligned with which words in French.
- $d(A|F, E)$ is a *distortion probability*. It indicates the probability of a particular order of French words, given the words and their English counterparts.

Distortion is often handled by a combination of heuristics and statistical estimates of very language-pair-specific word order transformations, e.g., French N_1 de $N_2 \longrightarrow N_2\ N_1$ in English.