PHRASE-BASED MODELS

- Current best performing SMT systems are not word-based, they are phrase-based: they are models that translate small sequences of words at a time.

- What is called "phrase" for these models is not linguistically motivated. It simply is a sequence of words.

MOTIVATIONS FOR PHRASE-BASED MODELS

1. WORDS MIGHT NOT BE THE BEST CANDIDATES AS SMALLEST UNITS OF TRANSLATIONS

One word in the source language translates into several words in the target language and vice versa.

<u>Natürlich</u> hat John <u>spass am</u> Spiel

<u>Of course,</u> John {has fun with the / enjoys} game / the game

+ Six German words and eight English words are best mapped by five translation units.

+ Notice the non-linguistic nature of "fun with the". Would it make better sense to translate

fun : spass and with the : am ?

It might make better segmentation sense from a linguistic point of view, but we would lose the linguistic and statistical information about the context.

<u>with the</u> is translated as <u>am</u> only in the context of <u>spass.</u>

MOTIVATIONS FOR PHRASE-BASED MODELS

+ TRANSLATING WORD SEQUENCES HELPS RESOLVE TRANSLATION AMBIGUITY

  "with the" is not a very common translation of "am", but in the context of "spass" it becomes the most common.

+ WORDS ARE NOT THE BEST MAXIMAL UNIT

  Phrases can be as long as one wants, memorising entire short sentences if necessary

+ PHRASE-BASED MODELS ARE CONCEPTUALLY SIMPLER

  Sequence of words in the source correspond to sequence of words in the target.

MATHEMATICAL DEFINITION OF A PHRASE-BASED MODEL

Recall that the final model will be a combination of a translation model and a language model.

$$\text{best}_e = \arg\max_e P(e|f)$$

$$= \arg\max_e \frac{P(f|e)\,P(e)}{\cancel{P(f)}}$$

language model will be the same as what we saw before

Translation model

$$P(f|e) = P(\bar{f}_1|\bar{e}_1)\,P(\bar{f}_2|\bar{e}_2)\ldots P(\bar{f}_n|\bar{e}_n)$$

So we can guarantee that the <u>number</u> of phrases will be the same in the two languages.?

But how do we know which are the corresponding f_i and e_i phrases?

MATHEMATICAL DEFINITION OF PBM (CONTINUED)

+ We don't know the alignement of the phrases, so we take all of them into account. (We will find a way of considering only those consistent with word alignment.)

+ In principle, we take all alignments into account, and we add a cost for translating phrases that are in a very different position in the source and in the target. This solution is not always correct (think of the position of German verbs) but it avoids excessive scrambling.

MATHEMATICAL DEFINITION OF PHRASE-BASED TRANSLATION

$$P(\bar{f}_1, \bar{f}_2, \bar{f}_3, \ldots, \bar{f}_I \mid \bar{e}_1, \bar{e}_2, \bar{e}_3, \ldots, \bar{e}_I) =$$

$$\prod_{i=1}^{I} t(\bar{f}_i \mid \bar{e}_i)\, d(start_i - end_{(i-1)} - 1)$$

+ t is a translation probability, but here it represents the probability of translating a sequence of words into another sequence of words, of any length.

+ d is a distortion parameter. It is based on the number of words skipped.
It is an exponentially decaying function.

DISTANCE-BASED REORDERING MODEL

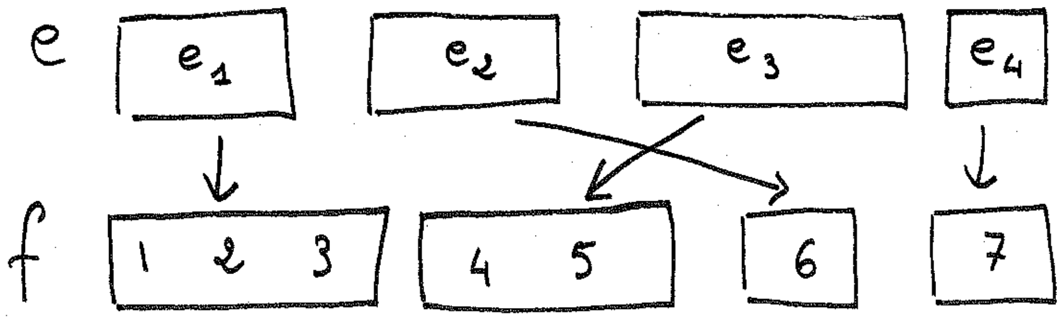
$$d(start_i - end_{i-1} - 1)$$

- $start_i$: position of the first word of the target input phrase that translates into the ith source phrase

- end_i : position of the last word of the target input phrase that translates into the ith source phrase.

Reordering distance is the number of words skipped (either backward or forward) when taking foreign words out of sequence.

DISTANCE-BASED REORDERING MODEL - EXAMPLE



$start_1 = 1 \qquad end_1 = 3$

$start_2 = 6 \qquad end_2 = 6$

$start_3 = 4 \qquad end_3 = 5$

$start_4 = 7 \qquad end_4 = 7$

$i=1 \quad d(start_1 - end_0 - 1) = d(1-0-1) = d(0)$

$i=2 \quad d(start_2 - end_1 - 1) = d(6-3-1) = d(2)$

$i=3 \quad d(start_3 - end_2 - 1) = d(4-6-1) = d(-3)$

$i=4 \quad d(start_4 - end_3 - 1) = d(7-5-1) = d(1)$

$d(x) = \alpha^{|x|}$ exponentially decaying cost function

LEARNING A PHRASE TRANSLATION TABLE

How do we learn a good translation table?

Two-step approach

1. create a word alignement between sentence pairs.
2. extract phrase pairs that are consistent with the word alignment.

+ We <u>create a word alignment</u> using EM and based on a simple word translation model.

+ We <u>extract phrase pairs</u> by
  - create a bidirectional alignment
  - grow alignment-consistent phrases.

CREATING A BIDIRECTIONAL ALIGNEMENT

- Recall that normal aligenement does not capture the fact that single words might be generated by more than one word.

  This is true in both directions.

  For example, let the following pair be given.

  Michael geht davon aus, dass er im Haus bleibt

  Michal assumes that he will stay in the house / at home

- In the direction from German to English, we have no way of indicating that <u>assumes</u> is generated by the triplet <u>geht davon aus</u>.

- In the direction from English to German then we have no way of indicating that <u>will stay</u> gives rise to <u>bleibt</u>.

- The solution is to generate the aligenement in both directions and then take the intersection (or the union, depending on whether we want good precision or good recall, respectively.)

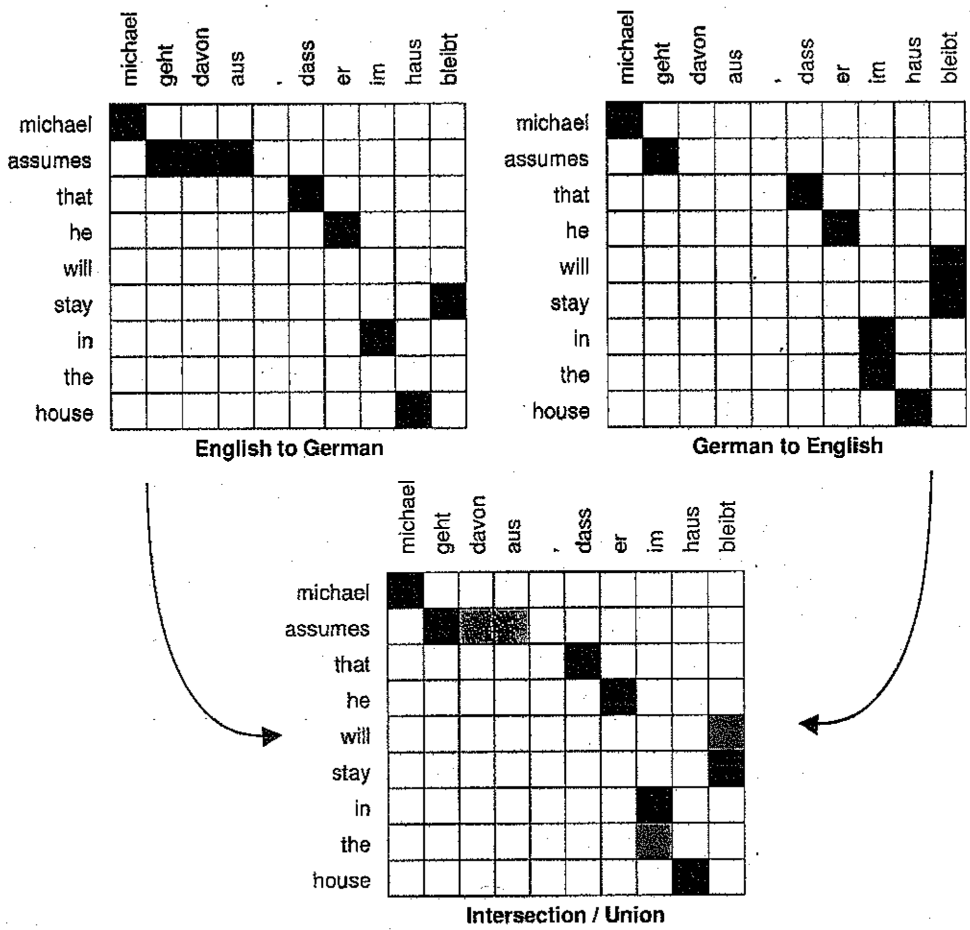
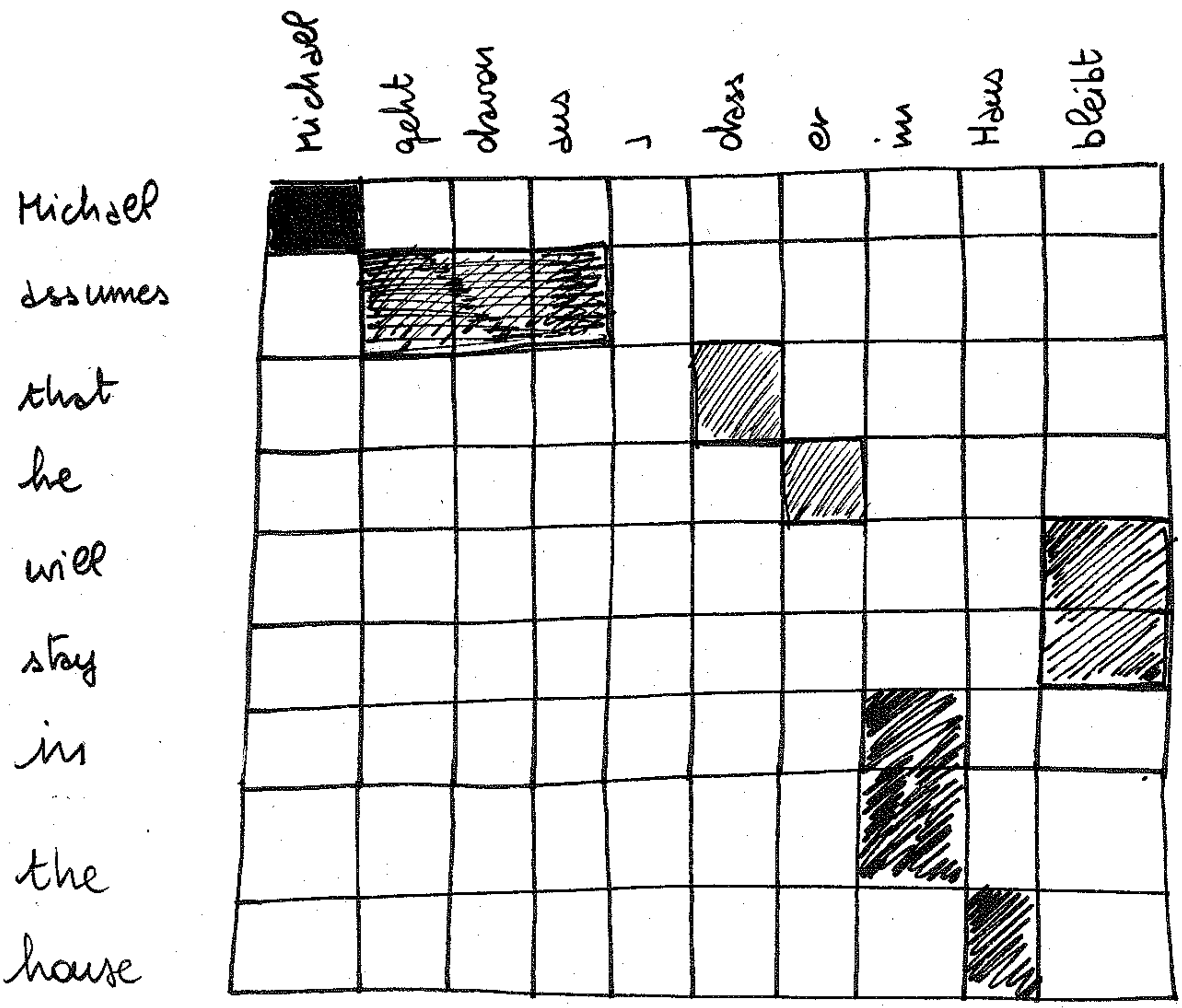


Figure 4.13: Symmetrization of IBM Model alignments. Since these models are not capable of aligning multiple input words to an output word, both a German–English and a English–German alignment will be faulty. However, these alignments can be merged by taking the intersection or union of the sets of alignment points.

CALCULATING THE SET OF CONSISTENT PHRASES



This is the alignment that results from the English-to-German and German-to-English.

Starting from this word alignment, we want to create a phrase alignment.

We want to extract both long and short phrases: short phrases are seen more frequently, so they help generalise, long phrases provide more context so they help disambiguate.

CONSISTENT PHRASE : DEFINITION

- A phrase pair $(\bar{f}, \bar{e})$ is consistent with a word aligmement A, iff all words $f_1 \dots f_n$ in $\bar{f}$ that have aligment points in A have these with words $e_1 \dots e_n$ in $\bar{e}$ and viceversa

- $(\bar{e}, \bar{f})$ is consistent with A iff

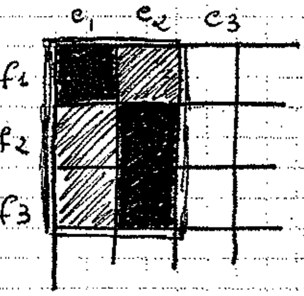
$$1.\ \forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f}$$

AND

$$2.\ \forall f_j \in \bar{f} : (e_i, f_j) \in A \rightarrow e_i \in \bar{e}$$

AND

$$3.\ \exists e_i \in \bar{e} ; \exists f_j \in \bar{f} : (e_i, f_j) \in A$$



This is a consistent phrase alignment.

1) $(e_1, f_1) \in A \rightarrow f_1 \in \bar{f}$ TRUE

$(e_1, f_2) \in A$ FALSE hence $f_2 \in \bar{f}$ TRUE

$(e_1, f_3) \in A$ FALSE hence $f_3 \in \bar{f}$ TRUE

$(e_2, f_1) \in A$ FALSE hence $f_1 \in \bar{f}$ TRUE

$(e_2, f_2) \in A \rightarrow f_2 \in \bar{f}$ TRUE

$(e_2, f_3) \in A \rightarrow f_3 \in \bar{f}$ TRUE

2) $(e_1, f_1) \in A \rightarrow e_1 \in \bar{e}$ TRUE

$(e_2, f_1) \in A$ FALSE

$(e_1, f_2) \in A$ FALSE

$(e_2, f_2) \in A \rightarrow e_2 \in \bar{e}$ TRUE

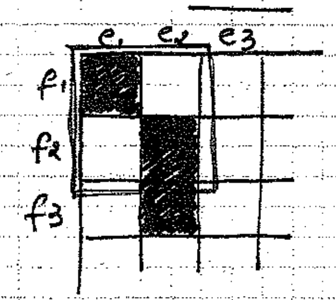
$(e_1, f_3) \in A$ FALSE

$(e_2, f_3) \in A \rightarrow e_2 \in \bar{e}$ TRUE

3) $(e_1, f_1) \in A$
$(e_2, f_2) \in A$
$(e_3, f_3) \in A$

This example is consistent because all words in the alignment are in the phrase alignments and so are all "projections" in the phrase

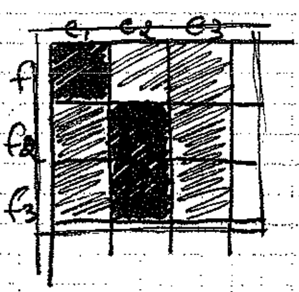
This is not a consistent phrase alignment



1) $\forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f}$ FALSE FOR

$(e_2, f_3) \in A \not\rightarrow f_3 \in \bar{f}$

N.B. This is a consistent phrase alignment



This is consistent even if it includes unaligned words because the last clause requires that all word aligned be included but not that only aligned words be included.

Unaligned words can belong to several alignments.

PHRASE EXTRACTION ALGORITHMS

Given the definition of consistent phrase alignement, what is the algorithm to extract consistent phrase pairs?

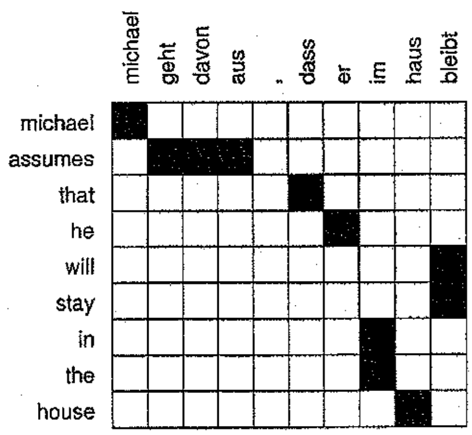
IDEA

- loop over all source language phrases and find the <u>minimal foreign phrase</u> matching.
- Matching is done as follows
  - find all the alignment points for the source phrase and find the shortest target phrase that includes all target counterparts for the source words.

  Constraints:
  - if no aligned source words, then no match
  - if matched target phrase has additional alignement points, then it is not consistent and it cannot be included.
  - if target phrase borders unaligned phrases, these are included and more than one match can be attributed to the original source phrase.

Phrase Extraction Algorithm - Example

This figure lists all the phrase pairs consistent with the alignment which will be extracted by the algorithm.



michael — michael
michael assumes — michael geht davon aus ; michael geht davon aus ,
michael assumes that — michael geht davon aus , dass
michael assumes that he — michael geht davon aus , dass er
michael assumes that he will stay in the house
— michael geht davon aus , dass er im haus bleibt
assumes — geht davon aus ; geht davon aus ,
assumes that — geht davon aus , dass
assumes that he — geht davon aus , dass er
assumes that he will stay in the house
— geht davon aus , dass er im haus bleibt
that — dass ; , dass
that he — dass er ; , dass er
that he will stay in the house
— dass er im haus bleibt ; , dass er im haus bleibt
he — er
he will stay in the house — er im haus bleibt
will stay — bleibt
will stay in the house — im haus bleibt
in the — im
in the house — im haus
house — haus

Figure 5.6: Extracted phrase pairs from the word alignment in Figure 5.3. For some English phrases, multiple mappings are extracted (e.g. *that* translates to *dass* with and without preceding comma), for some English phrases, no mappings can be found (e.g. *the* or *he will*).

- It is possible that, for some English phrase, the corresponding German phrase cannot be extracted.

  For example, he will stay aligns to er ... bleibt, but the intervening im Haus on the German side aligns to a phrase external to the English phrase he will stay, so it cannot be aligned.

- Unaligned words (the German comma) lead to multiple phrases extracted for the same English phrase.

- Allowing phrases of any length produces roughly a quadratic number of alignements, but very long phrases are usually not found in the training data, so usually a maximum phrase length is imposed.

ESTIMATING PHRASE TRANSLATION PROBABILITIES

- The estimation of the probability of translating phrases into other phrases is done by simple relative frequency, over all possible lengths of phrases (up to a bound that is determined for practical reasons).

For each sentence pair, a number of phrase pairs is extracted. We store the count of a particular phrase pair over all sentences, count$(\bar{e}, \bar{f})$.

The translation probability is then

$$t(\bar{f} \mid \bar{e}) = \frac{\text{count}(\bar{e}, \bar{f})}{\sum_{\bar{f_i}} \text{count}(\bar{e}, \bar{f_i})}$$