Introduction

Learning the Phrase Translation Table

Decoding

'Decoding' means trying to solve \( \arg\max_{\vec{s}, \tau} p(\vec{o}, \tau, \vec{s}) \), which is

\[
\arg\max_{\vec{s}, \tau} \left[ \prod_{k=1}^{K} \text{tr}(\vec{d}(\tau(k) | \vec{s}_k)) \ d(\vec{d}(\tau(k-1), \vec{s}_k)) \right] \ LM(s)
\]

- the decoding constructs \( \vec{s} \) in sequence \( \rightarrow \) need the score for a complete hypothesis \((\tau, \vec{s})\) to be computable from scores for partial ones \((\tau_k \vec{s}_{1:k})\)
- we still need to 'fold in' the \(LM\) contribution, but that is easily done. Let \(s_{1:i} | \text{let}(s_k)\) be the source sequence underlying the phrase sequence \(\vec{s}_{1:k}\), and \(\Phi_{LM}(\vec{s}_k)\) be the parts of the \(LM(s_{1:i} | \text{let}(s_k))\) mentioning the words in \(\vec{s}_k\). This is

\[
\Phi_{LM}(\vec{s}_k) = \prod_{i=\text{fst}(s_k)}^{i=\text{let}(s_k)} p(s_i | s_{i-\text{(n-1)}}, \ldots, s_{i-1})
\]

- the fully incremental equation is then

\[
\arg\max_{\vec{s}, \tau} \left[ \prod_{k=1}^{K} \text{tr}(\vec{d}(\tau(k) | \vec{s}_k)) \ d(\vec{d}(\tau(k-1), \vec{s}_k)) \Phi_{LM}(\vec{s}_k) \right]
\]
The Translation Options from the phrase-table

- For *er geht ja nicht nach hause* the above depicts a subset of the relevant phrase pairs (§
- in Europarl phrase table: 2727 matching phrase pairs for this sentence
- by pruning to the top 20 per phrase, 202 translation options remain
- picture shows top 4 for each § German ’phrase’

Exhaustive enumeration

- though its not realistic, its useful to first consider what would be involved in exhaustively enumerating all the possibilities, exploiting the incremental, LtoR nature of the defining formula

Exhaustive enumeration cntd

- initial hypothesis: no observed words used up, no source words produced; as observed words are used up will black out corresponding box in strip at top of box; will write in the box the chosen source phrase

Translation Options

- The correct translation chooses the highlighted phrase pairs above, and chooses an ordering
- has to somehow do this guided by the fact that $s_{post}$ is defined as the source part of

\[
\arg \max_{\bar{s}, \tau} \prod_{k=1}^{K} \phi(\delta_{r(k)}|\bar{s}_k) \cdot d(\delta_{r(k-1)}, \delta_{r(k)}) \cdot \Phi_{LM}(\bar{s}_k)
\]
Exhaustive enumeration contd

- put in all other hypotheses for first source phrase and matched observed phrase (only a few are shown)
- NOTE: all these partial hypotheses can be assigned a score as you go

Computational Complexity

- however, this outlined procedure is exponentially costly: for example if there are 2000 relevant entries from the translation tables for \(o = \text{er geht ja nicht nach hause}\), then there will already be approx 4,000,000 hypotheses for the first 2 phrases of the source.
- can actually prove that there is no non-exponential algorithm which is guaranteed to find the best translation as defined by formula (??) : 'Machine translation decoding is NP-complete'
- people use pruning in some fashion to kill the vast majority of hypotheses early on
- inherently risky as however you do it, you may prune something which looks insufficiently promising though it would recover as you go further
A different enumeration: by \( o \)-coverage

- In the above enumeration the hypotheses grew in stages, from all hypotheses about \( \bar{s}_{1:k} \) (the first \( k \) source phrases) to all hypotheses about \( \bar{s}_{1:k+1} \) (the first \( k + 1 \) source phrases). The simplest kind of pruning would set some max \( h_{\text{max}} \) and at each \( k \) keep only the best \( h_{\text{max}} \) hypotheses.
- This does not work very well though because the hypotheses concern wildly different numbers of source words and observed words, and so are not really comparable.
- An alternative is to change the book-keeping and progression a little so that all hypotheses that share the same amount \( L \) of coverage of observed words are grouped — say they are stored in \( \text{bin}_L \).
- **Generate**: if you have found all (or \( h_{\text{max}} \) best) hypotheses covering \( L \) observed words — \( \text{bin}_L \) — then can generate all successor hypotheses, adding each to relevant later \( \text{bin}_{L+1} \), where \( x \) is length of the chosen \( \bar{o} \) phrase of the successor hypothesis, and then move on to \( \text{bin}_{L+1} \).
- **Pruning**: can keep the size of such \( L \)-organised bins down by pruning, and in so doing be ranking more comparable things.

**er geht nach haus** example

To see this evolution with bins organised by \( o \)-coverage, suppose for *er geht nach haus* we have the following minute phrase table:

\[
\begin{array}{ccc}
\text{er} & \text{geht} & \text{nach} \\
\text{he} & \text{goes} & \text{home} \\
\end{array}
\]

Also without giving all the details of language model lets suppose the following source bigrams are impossible:

\[
\begin{array}{cc}
\text{he} & \text{home} \\
\text{goes} & \text{he} \\
\text{home} & \text{goes} \\
\end{array}
\]

The empty hypothesis is extended via possibilities in phrase table: creates some hypotheses in \( \text{bin}_1 \) and \( \text{bin}_2 \).
er geht nach haus: extend $\text{bin}_1$

Hypotheses in $\text{bin}_1$ now extended; creates additions to $\text{bin}_2$ and $\text{bin}_3$
impossible bigrams rule out
extension of *he* with *home*
extension of *goes* with *he*

er geht nach haus: extend $\text{bin}_2$

Hypotheses in $\text{bin}_2$ now extended; creates additions to $\text{bin}_3$ and $\text{bin}_4$
impossible bigrams rule out
extension of *home* with *goes*

er geht nach haus: extend $\text{bin}_3$

Hypotheses in $\text{bin}_3$ now extended; creates additions to $\text{bin}_4$
Tracing back from $\text{bin}_4$ there are 3 possible translations
*he goes home*; *goes home he*; *home he goes*
We did not include the scores from the model: these translations would be distinguished by the source lang. factor $LM$ and by the distortion factors, favouring *he goes home*

to get a further feel for what this alternative enumeration of the hypothesis space via amount of $\alpha$-coverage looks like the following pictures\textsuperscript{2} depict schematically an evolution of contents of bins; for the moment there is no pruning
to keep things compact just showing schematically the observed words consumed and not the associated chosen source phrase, nor the pointers back to the ancestor hypothesis
the example comes from a Chinese/English case. Though for the schematic pictures the actual translation options are not important, for completeness they are:

\textbf{bushi yu shalong juxing le huitan}

\begin{itemize}
  \item Bush
  \item and
  \item Sharon
  \item hold
  \item with
  \item held
  \item meeting
  \item juxing
  \item le
  \item huitan
  \item talks
  \item held a talk
\end{itemize}

\textsuperscript{2}From Watanabe
empty hypothesis, no observed words consumed

extend the empty $\text{bin}_0$ hypothesis, adding to $\text{bin}_1$ and later...

$\text{bin}_1$ is complete

extend $\text{bin}_1$ hypotheses, adding to $\text{bin}_2$ and later...
bin₂ is complete

extend bin₂ hypotheses, adding to bin₃ and later...

bin₃ is complete

extend bin₃ hypotheses, adding to bin₄ and later...
Beam search algorithm

With hypotheses stored in bins associated with number of $o$ words used, have a better basis for pruning and leads to the following simple so-called 'Beam search' version

place empty hypothesis into $b_{0}$

for $L = 0$ to $\text{length}(o) - 1$ {
    for each hypotheses in $b_{L}$ {
        for each applicable pair $\bar{s}, \bar{o}$ {
            create new hypothesis (including score)
            place in $b_{L+x}$ where $\bar{o}$ has length $x$
            prune $b_{L+x}$ if too big
        }
    }
}

why beam?

- 'beam': idea is that across the bins, which really should be exponentially growing, a fixed width 'beam' of good hypotheses is illuminated (ie. kept)
- its a heuristic and while it strives to have the $h_{\text{max}}$ options stored in $b_{L}$ be genuinely the $h_{\text{max}}$ best hypotheses that use $L$ observed words, it will not do this perfectly
- the width of the 'beam' $- h_{\text{max}}$ - is set as high as is computationally feasible and at the very end the best is chosen

following pictures show again an evolution of bins, this time including pruning
extend bin₀ hypotheses, adding to bin₁, …

prune bin₁

extend bin₁ hypotheses, adding to bin₂, …
prune \bin_2

extend \bin_2 hypotheses, adding to \bin_3, \ldots

prune \bin_3

extend \bin_3 hypotheses, adding to \bin_4, \ldots
prune bin₄