er geht ja nicht nach hause

are it goes does not go to home

yes he home
OVERVIEW

‣ Phrase based SMT
  ▶ Scoring formula
  ▶ Decoding algorithm

‣ Neural network ‘encoder-decoder’
WORD- AND PHRASE-BASED MT

- Seen word based models of translation
  - now used for *alignment*, but not actual *translation*
  - overly simplistic formulation

- Phrase based MT
  - treats n-grams as translation units, referred to as ‘phrases’ (not linguistic phrases though)

\[
\begin{array}{c}
\text{natuerlich} \\
\text{hat} \\
\text{john} \\
\text{spass am} \\
\text{spiel} \\
\text{of course} \\
\text{john} \\
\text{has} \\
\text{fun with the} \\
\text{game}
\end{array}
\]

Fig from Koehn09
PHRASE VS WORD BASED MT

- Phrase-pairs memorise:
  - common translation fragments (have access to local context in choosing lexical translation)
  - common reordering patterns (making up for naïve models of reordering)
“Extract” phrase pairs as contiguous chunks in word aligned text; then

- compute counts over the whole corpus
- normalise counts to produce ‘probabilities’

E.g.,

\[
\phi(\text{im haus bleibt}|\text{will stay in the house}) = \frac{c(\text{will stay in the house; im haus bleibt})}{c(\text{im haus bleibt})}
\]
The **phrase-table** consists of all phrase-pairs and their scores, which forms the search space for decoding

- E.g., for *natuerlich* it may contain the following translation phrases

| Translation | Probability $p(e|f)$ |
|-------------|---------------------|
| of course   | 0.5                 |
| naturally   | 0.3                 |
| of course , | 0.15                |
| , of course | 0.05                |

- Generally a massive list with many millions of phrase-pairs
DECODING

\[ E^*, A^* = \arg\max_{E, A} \text{score}(E, A, F) \]

- A describes the segmentation of \( F \) into phrases; and the re-ordering of their translations to produce \( E \)

- The score function is a product of the
  - translation “probability”, \( P(F|E) \), split into phrase-pairs
  - language model probability, \( P(E) \), over full sentence \( E \)
  - distortion cost, \( d(\text{start}_i, \text{end}_{i-1}) \), measuring amount of reordering between adjacent phrase-pairs

- Search problem
  - find translation \( E^* \) with the best overall score
Score the translations based on translation probabilities (step 2), reordering (step 3) and language model scores (steps 2 & 3).
Cover all source words exactly once; visited in any order; and with any segmentation into “phrases”

Choose a translation from phrase-table options

Leads to millions of possible translations…

Figure from Koehn, 2009
Akin to Viterbi algorithm

- factor out repeated computation (like Viterbi for HMMs, “chart” used in parsing)
- efficiently solve the maximisation problem

Aim is to translate every word of the input once

- searching over every segmentation into phrases;
- the translations of each phrase; and
- all possible ordering of the phrases
PHRASE-BASED DECODING

Start with empty state

Figure from Koehn, 2009
PHRASE-BASED DECODING

Figure from Koehn, 2009

Expand by choosing input span and generating translation
Consider all possible options to start the translation.
Continue to expand states, visiting uncovered words. Generating outputs left to right.
Read off translation from best complete derivation by back-tracking

Figure from Koehn, 2009
REPRESENTING TRANSLATION STATE

- Need to record
  - translation of phrase
  - which words are translated in bit-vector
  - last \( n-1 \) words in E… so that \( n \)-gram LM can compute probability of subsequent words
  - end position of the last phrase translated in the source, for scoring distortion in next step

- Together allows for the score computation to be factorised
COMPLEXITY

- Full search is intractable
  - word-based and phrase-based decoding is NP complete — arises from arbitrary reordering

- A solution is to prune the search space
  - Use \textit{beam search}, a form of approximate search
    - maintaining no more than \( k \) options (“hypotheses”)
    - pruning over translations that cover a given number of input words
PHRASE-BASED MT SUMMARY

- Start with sentence-aligned parallel text
  1. learn word alignments
  2. extract phrase-pairs from word alignments & normalise counts
  3. learn a language model

- Now decode test sentences using beam-search (where 2 & 3 above form part of scoring function)
Phrase-based approach is *rather* complicated!

Neural approach poses question:
- Can we throw away all this complexity, instead learn a single model to directly translate from source to target?

Using deep learning of neural networks
- learn robust representations of words and sentences
- attempts to generate words in the target given “deep” (vector/matrix) representation of the source
So-called “sequence2sequence” models combine:

- **encoder** which represents the source sentence as a vector or matrix of real values
  - akin to word2vec’s method for learning word vectors
- **decoder** which predicts the word sequence in the target
  - framed as a language model, albeit conditioned on the encoder representation
What is a vector representation of a sequence $\mathbf{x}$?
c = RNN(x)

What is the probability of a sequence $y \mid x$?
RNN ENCODER-DECODERS

\[ c = \text{RNN}(x) \]
\[ y \mid c \sim \text{RNNLM}(c) \]

What is the probability of a sequence \( y \mid x \)?
RNN ATTENTION MODEL

What is the probability of a sequence $y \mid x$?
What is the probability of a sequence $y$ | $x$?

Copyright 2018 The University of Melbourne

Slide credit: Duh, Dyer et al. 2015
What is the probability of a sequence $y \mid x$?
What is the probability of a sequence $y \mid x$?
What is the probability of a sequence $y | x$?
APPLICATIONS OF SEQ2SEQ

- Machine translation
- Summarisation (document as input)
- Speech recognition & speech synthesis
- Image captioning & image generation
- Word morphology (over characters)
  - e.g., study → student; receive → recipient; play → player; pay → payer/payee
- Generating source code from text & more...
EVALUATION: DID IT WORK?

- Given input in Persian

ملبورن مهد و مرکز پیش‌داشتن فنی مسازی و سیستم، تلویزیون، رقص باله، هنر امپرسیونیسم، سبک‌های مختلف رقص مثل دنگو و ملبورن شامل در استرالیا و مرکز مهم موزیک کلاسیک و امروزی در این کشور است.

- Google translate outputs the English

Melbourne cradle and center of origin of the film industry and cinema, television, ballet, art, impressionism, various dance styles such as New Vogue and the Melbourne Shuffle in Australia and an important center of classical and contemporary music in this country.

- Ask bilingual to judge? Ask to rate for two components

  - **fluency**: follows grammar of English, and semantically coherent
  
  - **adequacy**: contains the same information as the original source document

  - or edit the sentence until is is adequate, and measure #changes, time spent etc.
RESUABLE EVALUATION

- What if we have one (or several) good translations, e.g.

  Referred to as Australia’s “cultural capital” it is the birthplace of Australian impressionism, Australian rules football, the Australian film and television industries, and Australian contemporary dance such as the Melbourne Shuffle. It is recognised as a UNESCO City of Literature and a major centre for street art, music and theatre.

- We can use this text to evaluate many different MT system outputs for the same input
AUTOMATIC EVALUATION

‣ How many words are the shared between output:

Melbourne cradle and center of origin of the film industry and cinema, television, ballet, art, impressionism, various dance styles such as New Vogue and the Melbourne Shuffle in Australia and an important center of classical and contemporary music in this country.

‣ And the reference:

Referred to as Australia’s “cultural capital” it is the birthplace of Australian impressionism, Australian rules football, the Australian film and television industries, and Australian contemporary dance such as the Melbourne Shuffle. It is recognised as a UNESCO City of Literature and a major centre for street art, music and theatre.
MT EVALUATION: BLEU

- BLEU measures closeness of translation to one or more references
  - defined as:
    \[
    \text{BLEU} = \text{bp} \times \text{prec}_{1\text{-gram}} \times \text{prec}_{2\text{-gram}} \times \text{prec}_{3\text{-gram}} \times \text{prec}_{4\text{-gram}}
    \]
  - weighted average of 1, 2, 3 & 4-gram precisions
    - \( \text{prec}_{n\text{-gram}} = \frac{\text{num } n\text{-grams correct}}{\text{num } n\text{-grams predicted in output}} \)
    - numerator clipped to \#occurrences of \( n\text{-gram} \) in the reference
  - and a brevity penalty to hedge against short outputs
    - \( \text{bp} = \min (1, \frac{\text{output length}}{\text{reference length}}) \)

- Correlates with human judgements of fluency & adequacy
SUMMARY

- Word vs phrase based MT
  - Components of phrase-base approach
  - Decoding algorithm
- Neural encoder-decoder
- Evaluation using BLEU
- Reading
  - JM2 25.7 – 25.9
  - Neural Machine Translation and Sequence-to-sequence Models: A Tutorial, Neubig 201, Sections 7 & 8
    https://arxiv.org/abs/1703.01619