Multilingual and Parallel Corpora

Machine Translation (ctd)

Amir Zeldes

amir.zeldes@georgetown.edu

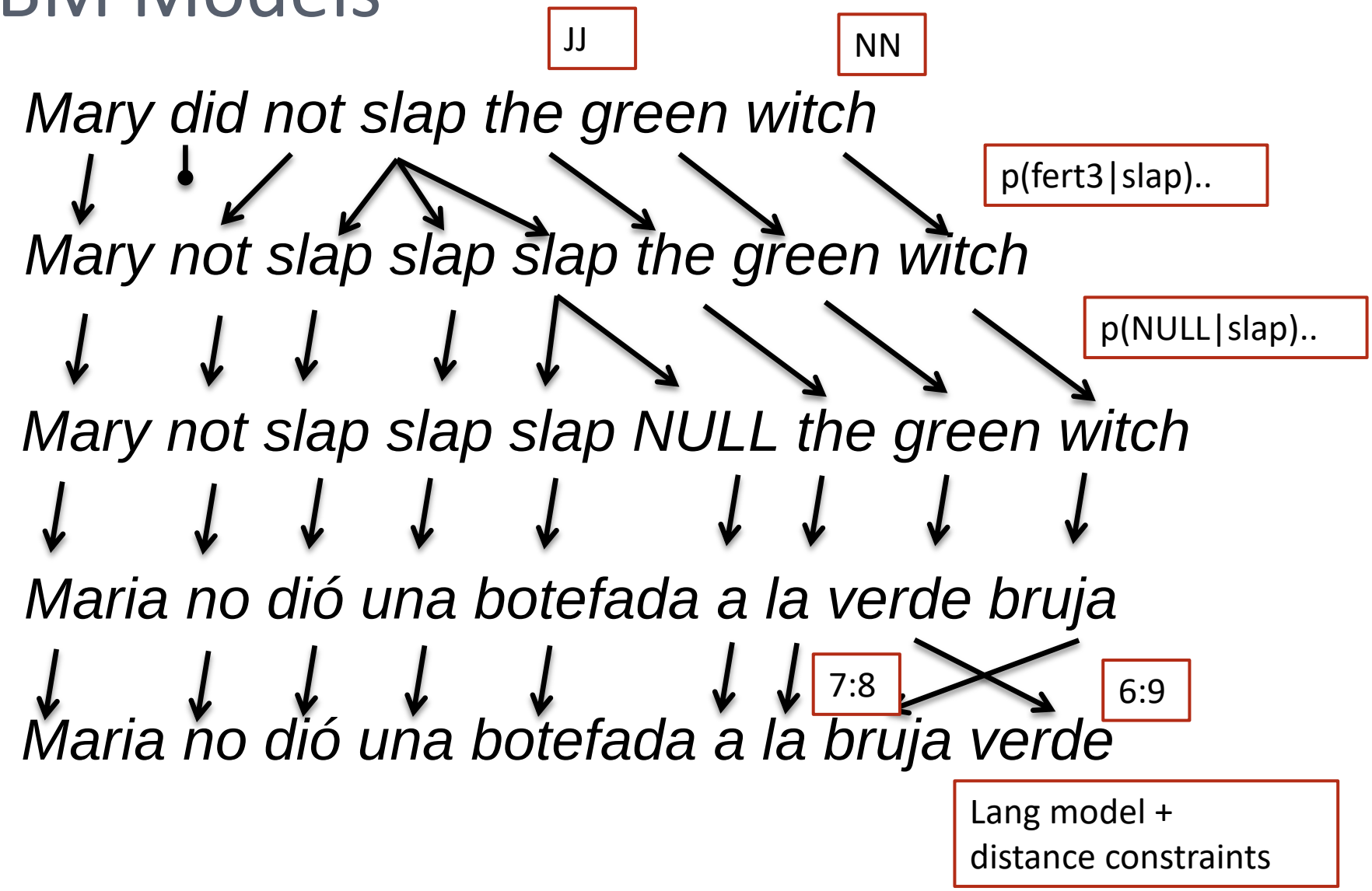
Noisy channel model

- Based on Bayes' Law, we can re-cast the MT problem:

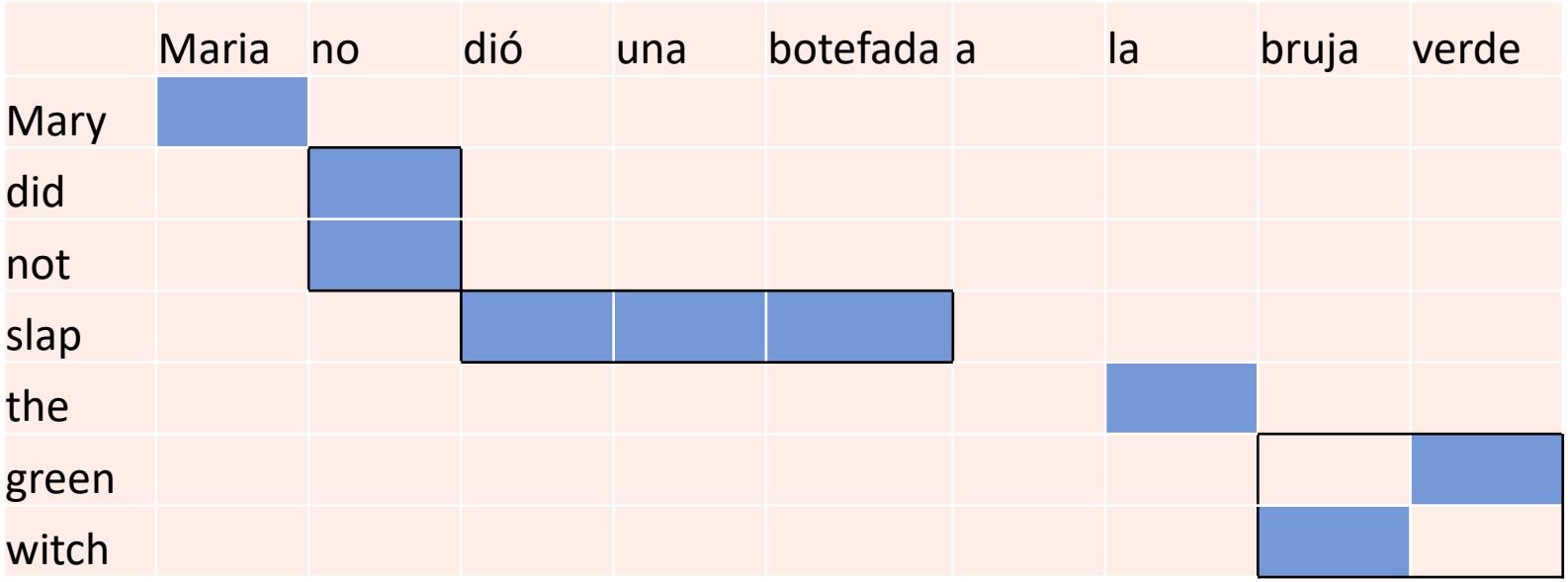
$$p(e|f) = \frac{p(e) \cdot p(f|e)}{p(f)}$$

- Separates the translation problem into two parts:
  - Finding translations
  - Assessing their *a priori* plausibility
  - Note that p(f) is a constant for any input sentence, so we can write: $p(e|f) = p(e) \times p(f|e)$

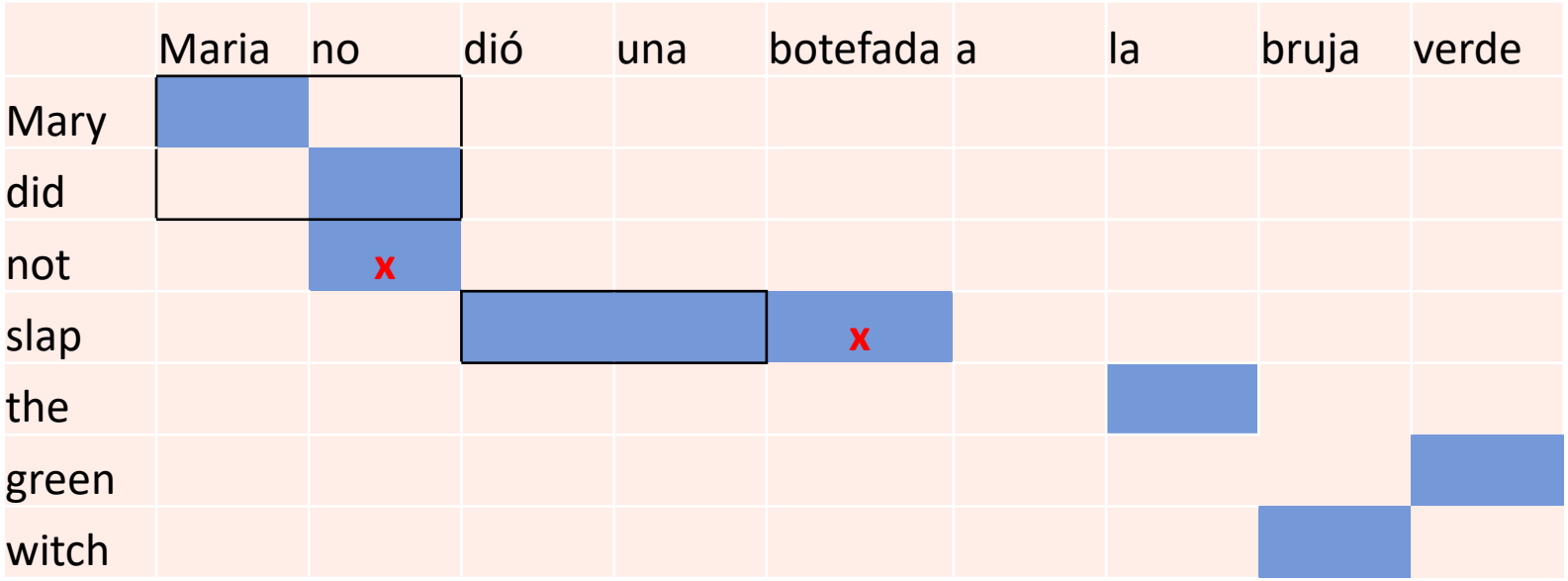
IBM Models


JJ
NN
Mary did not slap the green witch
p(fert3|slap)..
Mary not slap slap slap the green witch
p(NULL|slap)..
Mary not slap slap slap NULL the green witch
Maria no dió una botefada a la verde bruja
7:8
6:9
Maria no dió una botefada a la bruja verde
Lang model +
distance constraints

Automatic phrase alignment



Not aligned phrases



Phrase assumption - discussion

- Conceivably, individual language syntax is not the best representation for translation
- Are there cases in which we would like to learn translations for arbitrary chunks of language?

- Consider lexical bundles (Biber et al.):
  - *If you look at…*
  - *One of the most…*

Homework assignment – MT analysis

- **Error analysis** is crucial to developing MT further
  - Take a short text, up to about 200 words or a little less
  - Google Translate it to your L2
- Perform an error analysis:
  - Back-translate your translation manually to English
  - Place the side by side with the original
  - Mark differences
  - Classify differences as predominantly:
    - Translation model errors (conceivable L2, but wrong choice)
    - Language model errors (impossible in L2)

Example

- At this point, a movie studio would have to torch its headquarters, donate its merchandising revenues to charity, and produce a seven-hour art film performed in Ukrainian sign language to do something that truly qualified as a subversive gesture.

Example

- בשלב זה, אולפן סרטים יצטרך לפיד המטה שלה, לתרום הכנסות השיווק שלה לצדקה, ולהפיק סרט אמנות שבע שעות בצע בשפת סימנים אוקראינית לעשות משהו מוסמך באמת כמחווה חתרנית.
- At this stage, a movie studio would need a torch its headquarters/staff, donate some income her marketing to charity, and produce an art movie seven hours greed in Ukranian sign language to do something certified truly as a subversive gesture.

Differences

- At this point, a movie studio would have to torch its headquarters, donate its merchandising revenues to charity, and produce a seven-hour art film performed in Ukrainian sign language to do something that truly qualified as a subversive gesture.
- At this stage, a movie studio would need a torch its headquarters, donate some income her marketing to charity, and produce an art movie seven hours greed in Ukranian sign language to do something certified truly as a subversive gesture.

Analysis

- Translation model errors:
  - To torch -> לפיד - verb replaced by noun
  - Performed ->? בצע (greed) – totally inexplicable to me
  - Qualified -> מוסמך (certified) - potentially correct translation, but not in this context (wrong sense)

Analysis

- Language model errors:
  - הכנסות השיווק שלה - income her marketing – impossible Heb. sequence; coherence error with “her”, but definite object would require את“et”
  - Art movie seven hours – at a stretch, this could be understood as ‘produce a movie for seven hours’; a seven-hour movie would have to be סרט באורך שבע שעות “a movie at a length of seven hours”
  - …

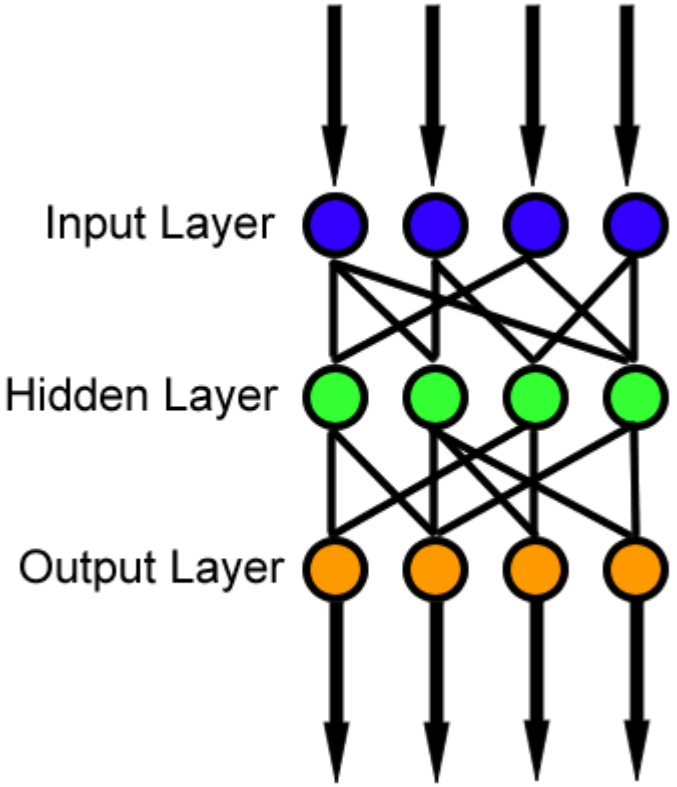
NMT – the state of the art

- Phrase based approaches were the state of the art until about 5 years ago
- New approach: Neural Machine Translation*
  - Reliance on machine learning to find best model
  - Deep Learning architectures allow special conditions to be learned for huge numbers of interacting features
  - Memory based architectures let the computer 'remember' having seen something to trigger a different translation

* This will be necessarily shallow – for more: https://bitbucket.org/hy-crossNLP/neuralmt/wiki/browse/

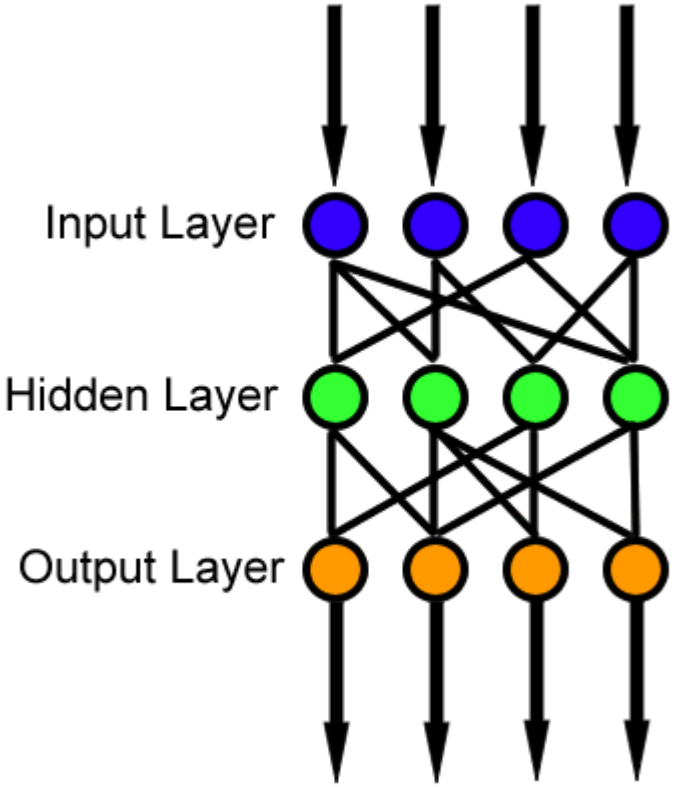
Feed forward networks

- Basic neural networks:
  - Take a bunch of inputs
  - Activate ‘synapses’
  - Propagate activation forward
  - Fire some output
- Input and output can be anything
- Word example:
  - Input: cat
  - Output: chat



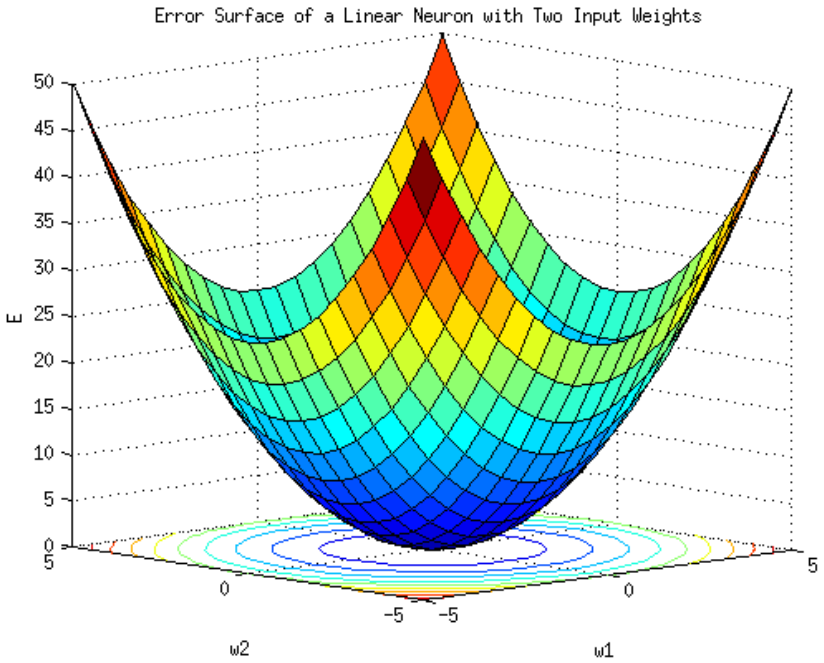
Back-propagation

- How do they learn?
  - Whenever the output is wrong we apply cost to all activating inputs
  - Propagate back in the network
  - Weights are modified
- Word example:
  - Input: cat
  - Output: chien -> all input links are penalized



Gradient descent

- How can we find the best weights?
  - Make fewest mistakes
  - Track derivative of cost (loss function)
  - If cost is getting less, all is well
  - If cost is getting higher, correct in other direction, until network converges on a minimal error

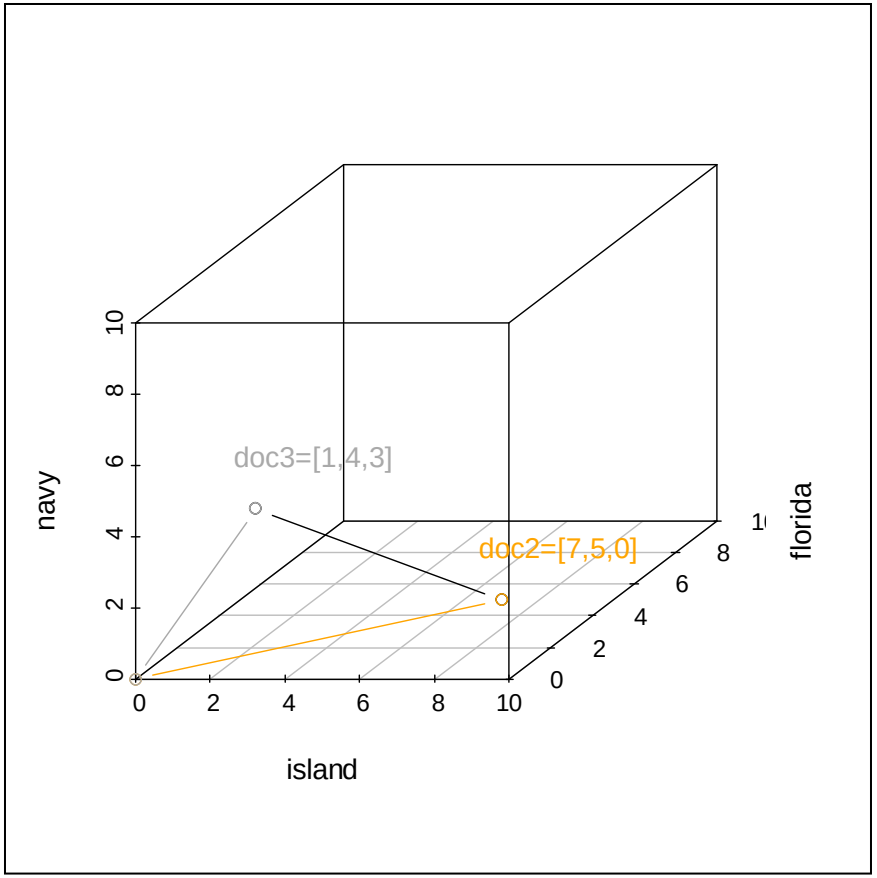


Is this a glorified lookup table?

- Although neural networks are very impressive, they are only as good as input/output features
  - Mapping words to words is not that amazing
  - Getting from words to sentences is still a problem
  - Many words will be out of data
- Real NMT approaches use many tricks to solve these problems

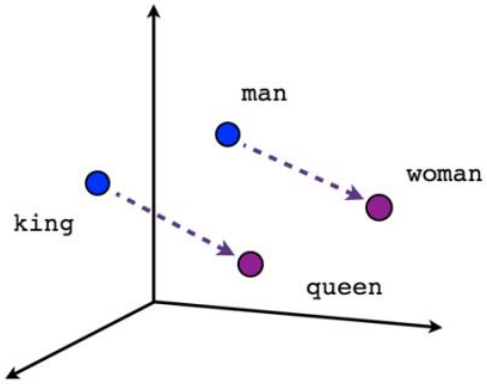
Vector space models / embeddings

- Lexical frequencies (or transformations thereof) make a vector space
- Allows similarity metrics for words and documents
- Models of meaning based on neighboring words

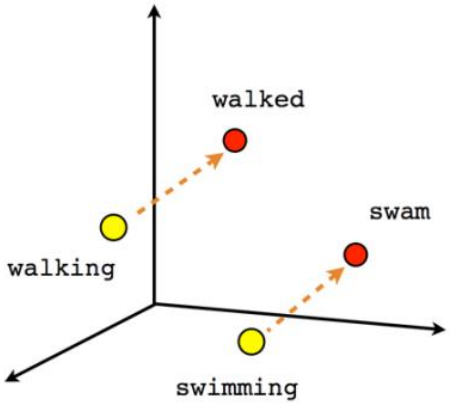


Applications

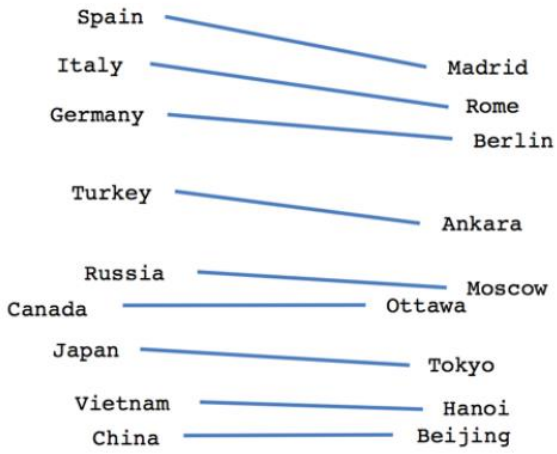
- Projecting vectors to lower dimensions
  - Reveal systematic relationships
  - Word level similarity



Male-Female



Verb tense



Country-Capital

Word2Vec

Do you think like vector space models?

- What is the most similar word to:
  - globe
  - sale
  - gum
- Analogy:
  - knife is to fork as chair is to…
  - water is to fish as air is to…
- Difference – which is the odd one out?
  - blue red green crimson transparent
  - computer table chair office

Try it: https://rare-technologies.com/word2vec-tutorial/

Multilingual models

- If we know some translations of neighbor words in SL and TL…
- We can train multilingual vector space models:
  - Similar coordinates not just for neighbors of government
  - Also for: *gouvernement*, *Regierung*…

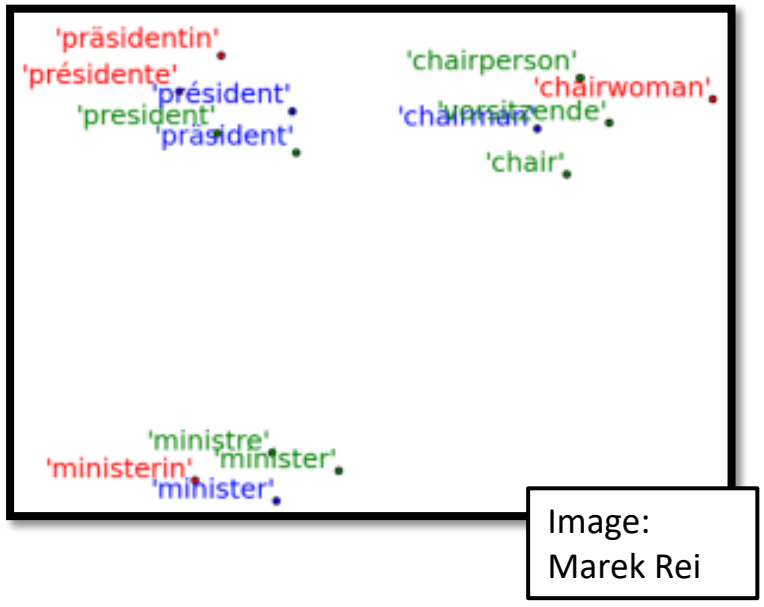
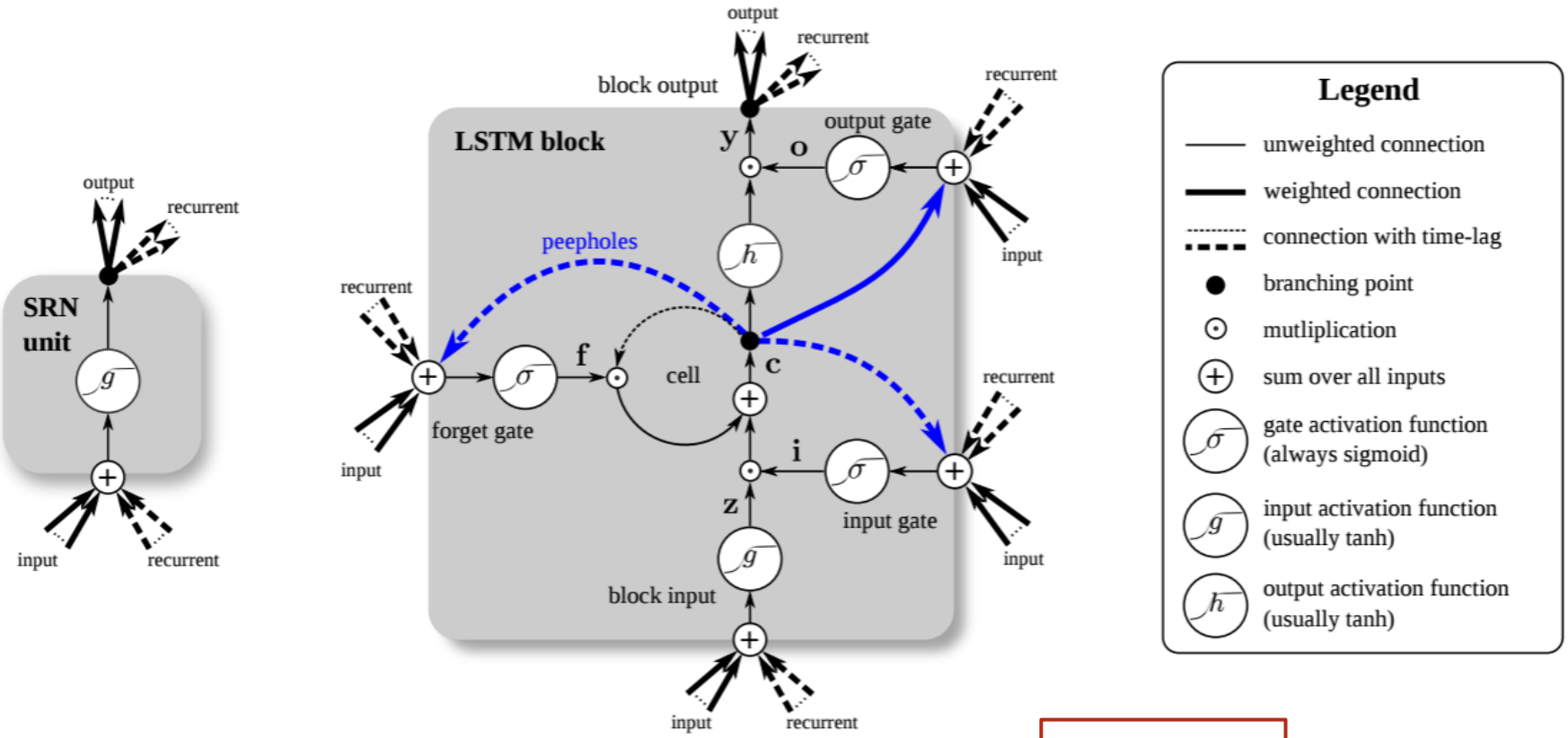


Image: Marek Rei

Using memory

- The latest NMT models also use memory based architectures (Recurrent Neural Networks [RNNs])
- Popular type: Long Short Term Memory (LSTM) networks
  - Cells don't just get input synapse weights, but also activate themselves
  - Allows cells to remember previous states
  - In LSTMs: cells also learn when to forget what they've seen

LSTM cell structure



Greff et al. (2015)

What can LSTMs do?

- Intuitive example: character level sequence to sequence modeling
- Example – trained on Shakespeare:

PANDARUS:
Alas, I think he shall be come approached and the day
When little srain would be attain'd into being never fed,
And who is but a chain and subjects of his death,
I should not sleep.

Second Senator:
They are away this miseries, produced upon my soul,
Breaking and strongly should be buried, when I perish
The earth and thoughts of many states.

DUKE VINCENTIO:
Well, your wit is in the care of side and that.

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

What can LSTMs do?

- Intuitive example: character level sequence to sequence modeling
- Example – trained on Wikipedia:

Naturalism and decision for the majority of Arab countries' capitalide was grounded by the Irish language by [[John Clair]], [[An Imperial Japanese Revolt]], associated with Guangzham's sovereignty. His generals were the powerful ruler of the Portugal in the [[Protestant Immineners]], which could be said to be directly in Cantonese Communication, which followed a ceremony and set inspired prison, training. The emperor travelled back to [[Antioch, Perth, October 25|21]] to note, the Kingdom of Costa Rica, unsuccessful fashioned the [[Thrales]], [[Cynth's Dajoard]], known in western [[Scotland]], near Italy to the conquest of India with the conflict. … Many governments recognize the military housing of the [[Civil Liberalization and Infantry Resolution 265 National Party in Hungary]], that is sympathetic to be to the [[Punjab Resolution]] (PJS)[http://www.humah.yahoo.com/guardian. cfm/7754800786d17551963s89.htm

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

What can LSTMs do?

- Example – trained on Linux source code:

```
/*
 * If this error is set, we will need anything right after
that BSD.
 */
static void action_new_function(struct s_stat_info *wb)
{
  unsigned long flags;
  int lel_idx_bit = e->edd, *sys & ~((unsigned long)
*FIRST_COMPAT);
  buf[0] = 0xFFFFFFFF & (bit << 4);
  min(inc, slist->bytes);
  printk(KERN_WARNING "Memory allocated %02x/%02x, "
    "original MLL instead\n"),
    min(min(multi_run - s->len, max) * num_data_in),
    frame_pos, sz + first_seg);
  return disassemble(info->pending_bh);
}

static void num_serial_settings(struct tty_struct *tty)
{
  if (tty == tty)
    disable_single_st_p(dev);
  pci disable_spool(port);
  return 0;
}
```

http://karpathy.github.io/2015/05/21/rnn-effectiveness/

What are these cells learning?

Cell sensitive to position in line:

```
The sole importance of the crossing of the Berezina lies in the fact
that it plainly and indubitably proved the fallacy of all the plans for
cutting off the enemy's retreat and the soundness of the only possible
line of action--the one Kutuzov and the general mass of the army
demanded--namely, simply to follow the enemy up. The French crowd fled
at a continually increasing speed and all its energy was directed to
reaching its goal. It fled like a wounded animal and it was impossible
to block its path. This was shown not so much by the arrangements it
made for crossing as by what took place at the bridges. When the bridges
broke down, unarmed soldiers, people from Moscow and women with children
who were with the French transport, all--carried on by vis inertiae--
pressed forward into boats and into the ice-covered water and did not,
surrender.
```

Cell that turns on inside quotes:

```
"You mean to imply that I have nothing to eat out of.... On the
contrary, I can supply you with everything even if you want to give
dinner parties," warmly replied Chichagov, who tried by every word he
spoke to prove his own rectitude and therefore imagined Kutuzov to be
animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating
smile: "I meant merely to say what I said."
```

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
        siginfo_t *info)
{
  int sig = next_signal(pending, mask);
  if (sig) {
    if (current->notifier) {
      if (sigismember(current->notifier_mask, sig)) {
        if (!(current->notifier)(current->notifier_data)) {
          clear_thread_flag(TIF_SIGPENDING);
          return 0;
        }
      }
    }
    collect_signal(sig, pending, info);
  }
  return sig;
}
```

A large portion of cells are not easily interpretable. Here is a typical example:

```
/*  Unpack a filter field's string representation from user-space
 *  buffer.  */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
  char *str;
  if (!*bufp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /*  Of the currently implemented string fields, PATH_MAX
   *  defines the longest valid length.
   */
```

Bonus fun

- You can chat with a neural network trained on conversational pairs
- Example:
  - http://neuralconvo.huggingface.co/
  - (also compare Microsoft's TAY: https://twitter.com/tayandyou)