# A recurrent neural network based statistical machine translation system

Hieu H. Pham Christopher D. Manning
hyhieu@cs.stanford.edu

CURIS Project, Summer 2014

# Motivations

From machine translation perspective

- MT systems are complicated
  - (NO offensive, but...) think about Stanford Phrasal MT.
  - Word/phrase alignment, LM, rule extraction, phrase table, decoder, tuning with large datasets, constituency,...
- For some languages (Hi, Zh, Ru, ...), data are scarce.

| File | Size | CS-EN | DE-EN | HI-EN | FR-EN | RU-EN |
|---|---|---|---|---|---|---|
| Europarl v7 | 628MB | ✓ | ✓ | | ✓ | |
| Common Crawl corpus | 876MB | ✓ | ✓ | | ✓ | ✓ |
| UN corpus | 2.3GB | | | | ✓ | |
| News Commentary | 77MB | ✓ | ✓ | | ✓ | ✓ |
| $10^9$French-English corpus | 2.3 GB | | | | ✓ | |
| CzEng 1.0 | 115MB | ✓ | | | | |
| Yandex 1M corpus | 121MB | | | | | ✓ |
| Wiki Headlines | 7.8MB | | | ✓ | | ✓ |
| HindEnCorp | 25MB | | | ✓ | | |
| ~~The JHU Corpus~~ | | | | ~~✓~~ | | |

Figure : Parallel data from WMT 2014

# Motivations

From language model (LM) perspective

- Recurrent neural network (RNN) can capture information about variable length sequences, which can later be decoded.
- LMs can generate *meaningful* text sequences

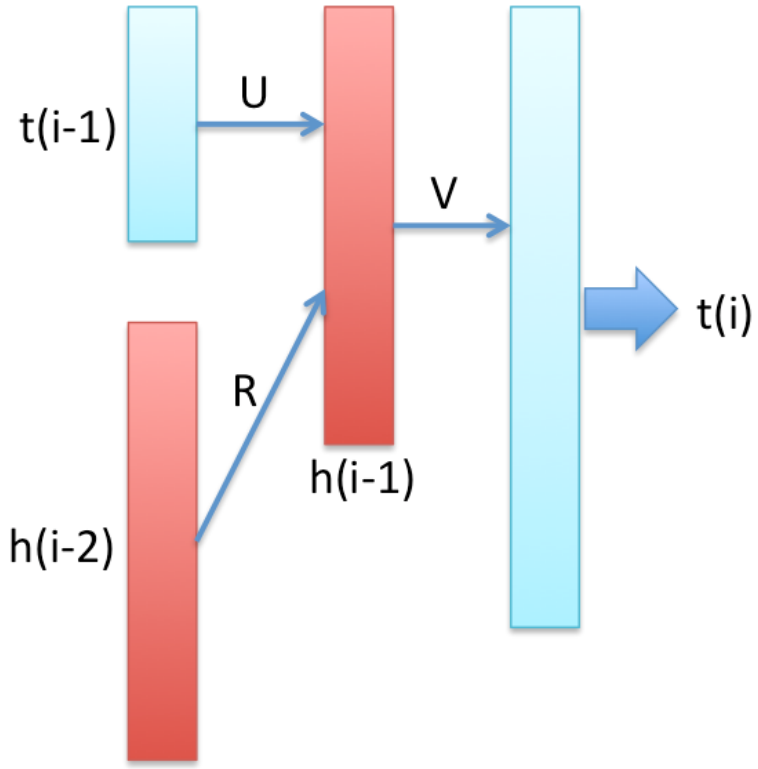
```
OR STUDENT'S IS FROM TEETH PROSECUTORS DO FILLED WITH
HER SOME BACKGROUND ON WHAT WAS GOING ON HERE
ALUMINUM CANS OF PEACE

THANKS FOR COMING IN NEXT IN A COUPLE OF MINUTES
WHEN WE TAKE A LOOK AT OUR ACCOMPANYING STORY IMAGE
GUIDE WHY ARE ANY OF THOSE DETAILS BEING HEARD IN LONDON
BUT DEFENSE ATTORNEYS SAY THEY THOUGHT THE CONTACT WAS
NOT AIMED DAMAGING AT ANY SUSPECTS

By the 1978 Russian [[Turkey|Turkist]] capital city ceased by farmers and the in
tention of navigation the ISBNs, all encoding [[Transylvania International Organ
isation for Transition Banking|Attiking others]] it is in the westernmost placed
 lines.  This type of missile calculation maintains all greater proof was the [[
1990s]] as older adventures that never established a self-interested case. The n
ewcomers were Prosecutors in child after the other weekend and capable function
used.
```

Figure : Texts generated by some language models. Upper: 5-gram (Schwenk et al., 2007); Middle: RnnLM (Mikolov et al., 2011); Lower: Character based Deep RNN (Graves, 2014).

# Recurrent neural network language model (RnnLM)




$$h_{i-1} = f\left(U \cdot t_{i-1} + R \cdot h_{i-2}\right)$$

$$p\left(t_i = w | t_{1..i-1}\right) = g\left(V \cdot h_{i-1}\right)$$

$$f(x) = \frac{1}{1+e^{-x}}, g(x) = \frac{e^x}{\sum_j e^{x_j}}$$

Figure : Standard RnnLM (Mikolov et al., 2011)

# Joint bilingual RnnLM

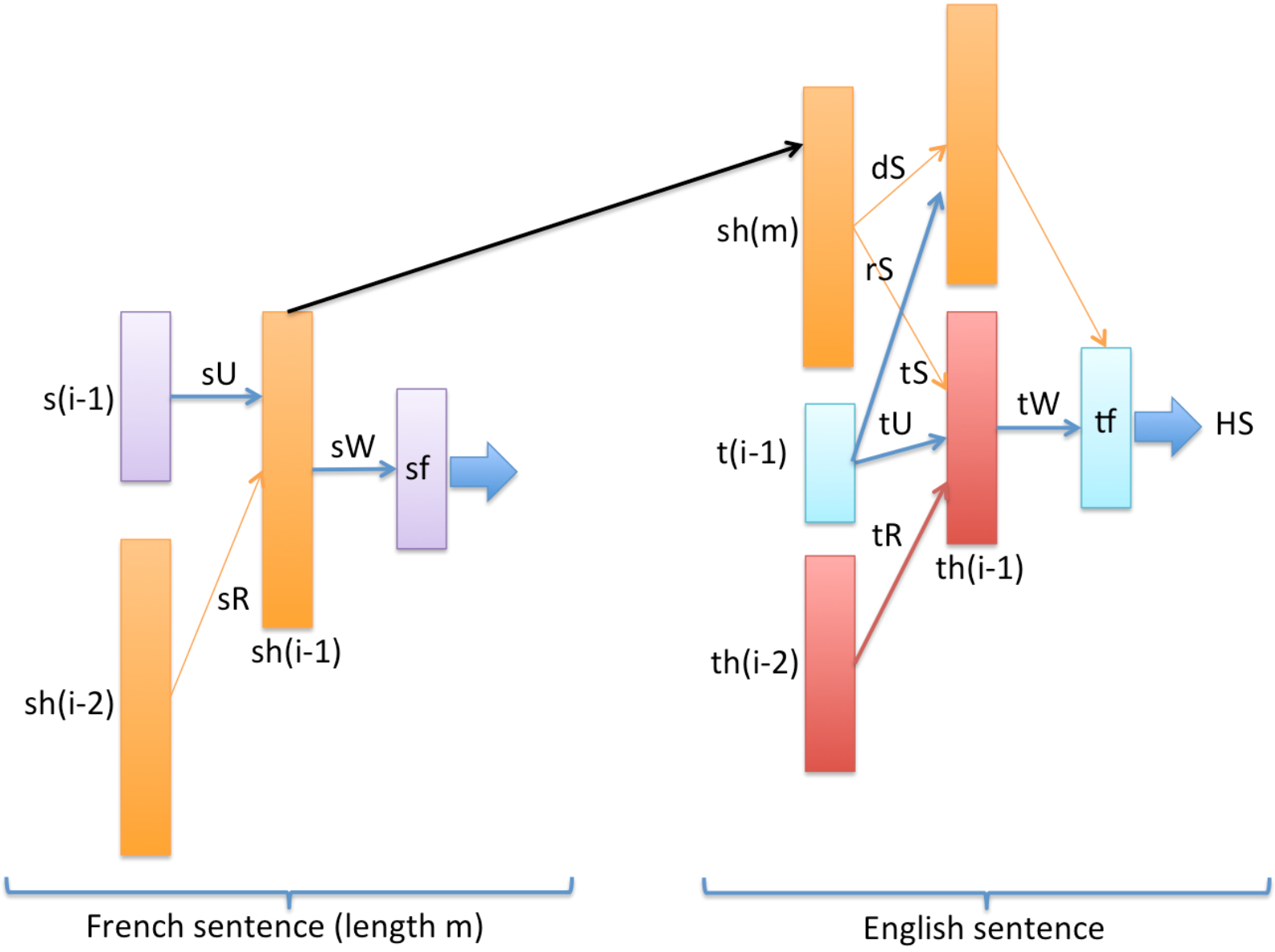



Figure : Architecture of source unrolled joint bilingual RnnLM.

# Joint bilingual RnnLM

- Source (fr): $s_1 s_2 \cdots s_m$; target (en): $t_1 t_2 \cdots t_n$
- Hypotheses:
  - The hidden layer of RnnLM can capture important semantics meaning of partial histories of the sentence
    - Thus the final values of the hidden layer capture the whole sentence's meaning.
  - These information can be "retrieved"
- We use the "meaning" of the source to bias the RnnLM of the target

$$tgtH_{i-1} = f\left(U \cdot t_{i-1} + R \cdot tgtH_{i-2}\right)$$

$$\text{becomes } tgtH_{i-1} = f\left(U \cdot t_{i-1} + R \cdot h_{i-2} + S \cdot srcH(m)\right)$$

- This can be extended further to "guide" the target model to "unroll" $srcH(m)$ and retrieve its meaning.

# Experiments

- Data: Europarl v7 fr-en, WMT '14
  - $2.2M$ parallel sentences, leaving out $20K$ sentences for validation.
  - Vocab sizes: $|V_{En}| = 121735$, $|V_{Fr}| = 138467$.
- Training is done via stochastic gradient descent
  - Minimize sum of cross entropy errors at each word in both source and target sentences.
  - 7 epochs with learning rate decaying after each epoch.
  - Gradient is computed via the traditional backprogation through time (BPTT) algorithm.
    - We did not truncate BPTT, nor did we clip the gradient to lower than some value.
    - We observed no gradient explosion as reported and cautioned in (Mikolov et al., 2011), (Le et al., 2010).
- The learned model are then used to generate "translations" via a beam decoder for best target sequences (not yet...)

# Results

- Perplexities

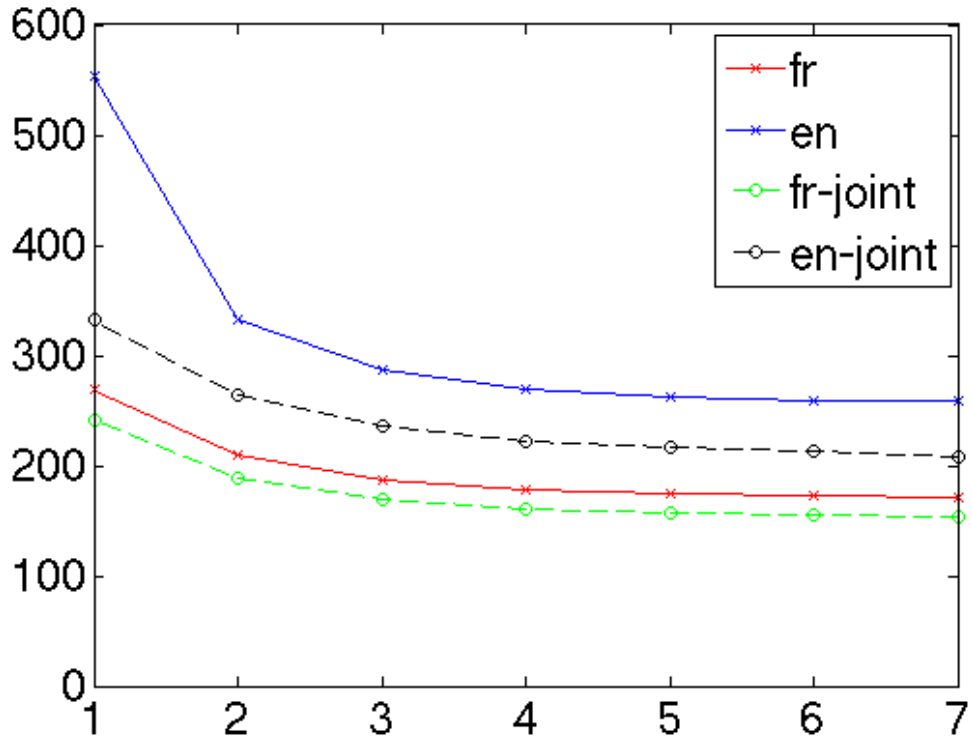



Figure : Perplexities on joint model are smaller than those on independent models.

# Results

- Lexical translation
  - Given a French word *f*, we encode it using the RNN model on source side, then compute the probabilistic distribution over the words in target dictionary.
  - Translated words are compared against the first word translated by Google Translator.
  - Result: 81.3% lexical translations match!
    - We have (nearly) inferred the dictionary!
  - Some wrong translations

| French | English gold | English translated |
|---|---|---|
| legalement | legally | legal |
| meme | same | , (comma) |
| exactement | exactly | right |
| connatre | know | taste |
| confidentialit | confidentiality | ¡unk¿ |


Thank You