Domain Adaptation via Biased Sampling

Ulrich Germann
Marcello Federico
Bruno Pouliquen
Huei-Chi Lin
Amittai Axelrod
Ali Hosseinzadeh Vahid

Progress Report

Introduction

- ***Sampling phrase tables*** compute phrase table entries on the fly by looking at a number of phrase occurrences.
- Currently, all samples are equally likely to be picked.
- Can we get better translations if we bias the sampling to prefer phrase occurrences in documents in the training corpus that are similar to the translation job?
- What's the best way to define similarity for this purpose?

Current state of the project

- ✓ got everyone on the same page
- ✓ decided on a data set to use (TED talks, en->fr)
- ✓ set up training/dev/test corpus
- ✓ developed ideas to measure similarity
- ✓ built baseline system
- ✓ implemented various document similarity measures
- ☹ couldn’t tune and evaluate yet due to technical problems

Preparation

IWSLT 2014 English-French Benchmark

train: TED Talk collection (1415 talks)
dev: dev2010 (8 talks), tst2010 (11 talks)
test: tst2011 (8 talks) tst2012 (11 talks)

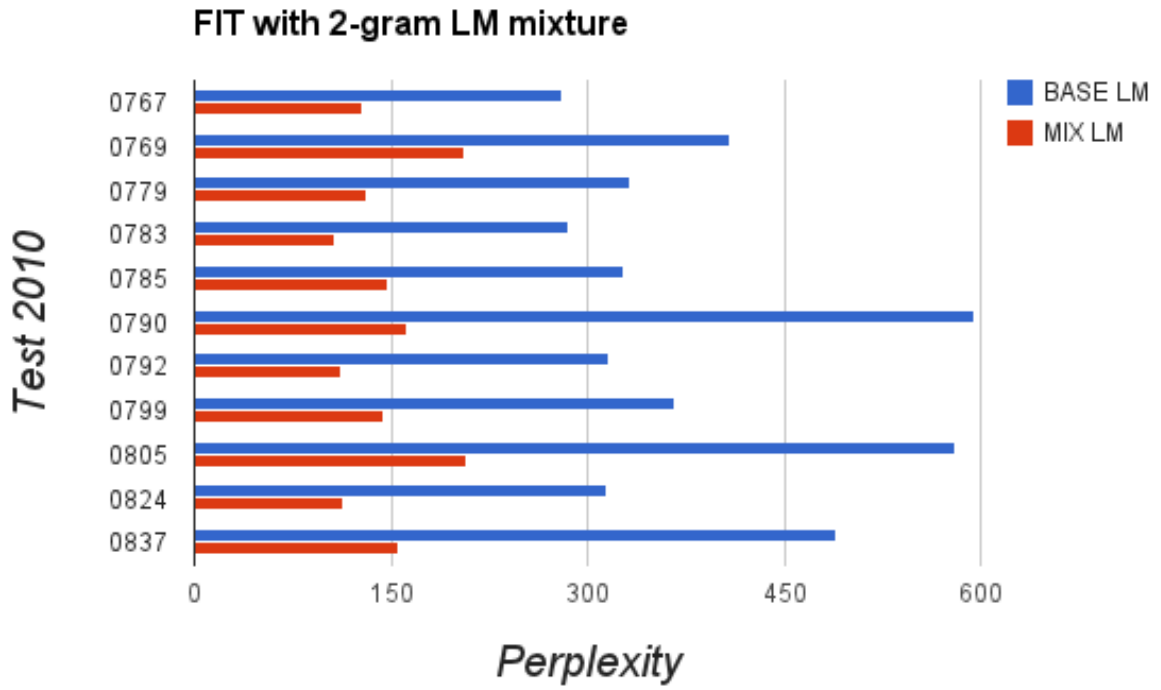
Preparation

- parallel data with word alignments
- English data with **talk id** at sentence level
- French language model (only in-domain data)
  - 5gr LM Improved Kneser-Ney (no pruning)
  - tst2012: PP=89 OOV < 1%
- English data with POS tags (Stanford tagger, v3.4.1)
- training SMT system

Idea 1: n-gram similarity

- Measure similarity between test and training talks
- Train word-based 2-gram LMs for each train talk
- Create a mixture of 1415 LMs !!
- For each test talk
  - estimate mixture weights with EM
  - use weights as optimal doc distribution

Idea 1: n-gram similarity



Idea 1: n-gram similarity

Computing PP on 0767
%% Nw=5620 PP=113.38 PPwp=5.39 Nbo=540 Noov=17 OOV=0.30%
%% Nw=6117 PP=102.94 PPwp=7.82 Nbo=543 Noov=30 OOV=0.49%

Uniform mixture for 0767
%% Nw=5620 PP=161.30 Nbo=540 Noov=17 OOV=0.30% Noov_any=5012 OOV_any=89.18%
%% Nw=6117 PP=160.06 Nbo=543 Noov=30 OOV=0.49% Noov_any=5643 OOV_any=92.25%

Training mixture for 0767 (on source)
Nw=5620 PP=127.72 Nbo=540 Noov=17 OOV=0.30% Noov_any=5012 OOV_any=89.18%
Nw=6117 PP=140.76 Nbo=543 Noov=30 OOV=0.49% Noov_any=5643 OOV_any=92.25%

Idea 2: semantic similarity (MF)

- PLSA topic model on talks
- Get talk-topic distribution of train data
- Infer topic distribution of each test talks

Still to be done….

Idea 3: Similarity

Create an index for the trainset (Lucene, Terrier)

Query the index (TFIDF, BM25…) with each document from the testset

Compute a (normalized) similarity score for each document

Stemming? Stopwords?

Idea 4: syntactic similarity

- compute sequences of POS tags found in the training data and dev set;
- compare their frequencies;

Idea 5: Discriminative, Style-based

Identify and keep words that:

* discriminate between talks in TED:
  high IDF / occur in at most 20% of the talks
* are frequent enough to matter:
  appear at least 10 times

Total of 11,415 words (out of 54,732).

Replace all other words with their POS tags:
information rich, and robust statistics.

Idea 5: Discriminative, Style-based

For each talk to be translated, we:

* Built vector of empirical frequencies of the 11k words + POS tags within the talk.
* Same for each talk in the training set
* Computed cosine similarity between target talk and training talks.
* Ranked training talks, fed to Uli.

Results

Trained on 1415 TED talks only (LM, TM):

- BLEU scores between 30.x and 37.x depending on test set for the baseline system

- couldn’t get tuning and eval working due to technical difficulty