

Confidence Estimation for Statistical Machine Translation

George Foster–U de M, Simona Gandrabur–U de M,
Cyril Goutte–XRCE, Lidia Mangu–IBM,
Erin Fitzgerald–JHU, Alberto Sanchis–UPV, Nicola Ueffing–RWTH,
John Blatz du Rivage–Princeton, Alex Kulesza–Harvard

Confidence Estimation

- CE is the attempt to determine whether NLP output is correct or not in a given context
- useful for practical applications of imperfect NLP technologies
- extensive previous work in SR, eg spoken dialog systems, but almost no work outside SR. . .
- motivation for workshop: apply CE to another area of NLP; assess performance and attempt to draw general conclusions

What is CE?

Goal: characterize the behaviour of some base NLP system that computes a function $y = \text{NLP}(x)$

Two versions of CE:

- **weak CE**: make correctness judgements on NLP output using a scoring function $S(\text{NLP}, x, y)$ and a threshold τ :

$$C(\text{NLP}, x, y) = \begin{cases} 1, & S(\text{NLP}, x, y) \geq \tau \\ 0, & \text{else} \end{cases}$$

- **strong CE**: estimate probabilities of correctness:

$$S(\text{NLP}, x, y) = p(C = 1 | \text{NLP}, x, y)$$

Why do CE?

Many applications:

- improved user interaction—use confidence estimates to minimize human effort; eg:
 - SR dialog: seek confirmation only when really needed
 - MT postediting: focus revisor's attention on problem areas
- bootstrapping (supervised) learning: user corrects selected examples in NLP output
- model combination: strong CE provides a uniform framework for combining output from many different base NLP systems (possibly non-probabilistic ones)
- recombination/search: for sequence output, use sub-sequence confidence scores to splice together parts of an nbest list

Methods

Two main approaches:

- use scores or probabilities derived from the base NLP system
 - simple and often effective (Wessel et al 2001)
 - need to smooth well, and do careful posterior probability estimates
- use a separate CE layer
 - ML techniques to learn NLP performance on an annotated validation corpus, correcting for training corpus error
 - modularity advantages:
 - * separate problems of picking best solution from determining whether it's correct: different features may be applicable, eg source sentence length in MT

- use a separate CE layer (continued)
 - advantages due to modularity (continued):
 - * may be possible to reuse CE techniques for different base NLP systems
 - * CE layer is smaller and easier to adapt to new domains than the base NLP system

CE for SMT Workshop Overview

- base SMT system(s) for Chinese to English translation:
 - alignment template (AT) (Och et al. 2003)
 - CMU SMT system (?)
 - data: nbest lists for NIST eval corpora (approx 2K source sentences) + word alignments + reference translations
- estimate various probabilities of correctness from this data, evaluate using abstract measures *and* within applications
- assess the value of different features
- compare different ML methods and configurations (NN's versus SVM's versus none; classification versus regression)
- challenges: sparse but large data, low SMT performance, many right answers → MT evaluation is hard

Sentence-level CE

- main problem: correctly translated sentences are rare \Rightarrow redefine correctness as having error measure (WER, -NIST, -BLEU) below a certain threshold
- ML approaches:
 - regression: learn error functions directly, apply threshold after
 - classification: tag = 0/1 as determined by the error threshold
- applications:
 - model combination AT + CMU
 - filtering for postediting simulation
 - bootstrap learning: eliminate noise from the training corpus
 - reranking

Sub-sentence CE

- estimate correctness probabilities for partial translations: words and low-order ngrams
- condition on: position, aligned source region, nothing
- problems:
 - sparseness in nbest lists \Rightarrow seek equivalence classes
 - probability may not reflect fit in current context, given existence of multiple correct translations, eg:

a b a

b b c

c b b

b a b

Sub-sentence CE (2)

Applications:

- identify error regions for postediting
- hypothesis recombination

Alignment CE

- estimate correctness for alignment links
- problem: don't have reference alignments
- application: any application for word-aligned data, eg MT

Confidence Estimation for Statistical Machine Translation

George Foster, RALI, Université de Montréal

Simona Gandrabur, RALI, Université de Montréal

`{foster, gandrabu}@iro.umontreal.ca`

`http://www-rali.iro.umontreal.ca`

George Foster–U de M, Simona Gandrabur–U de M,

Cyril Goutte–XRCE, Lidia Mangu–IBM,

Erin Fitzgerald–JHU, Alberto Sanchis–UPV, Nicola Ueffing–RWTH,

John Blatz du Rivage–Princeton, Alex Kulesza–Harvard



CLSP WS03
