Comparison of MT between related and unrelated languages

Ondřej Bojar, Natalia Kjlueva, David Kolovratník

Charles University in Prague,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

September 27, 2009

http://ufallab2.ms.mff.cuni.cz/~bojar/teaching/NPFL087/wiki/CzeRu
machine translation

in our experiment

▶ a system of programs
▶ takes text (natural language) as input, also needs models
▶ outputs text translated into another language
▶ poor quality – Does it worth reading?

different approaches

▶ data driven
  ▶ word to word
  ▶ phrase based
    ▶ example based, employing syntax, …
▶ manually constructed translation rules, …
phrase based machine translation – simplified idea

training/learning

▶ explores parallel bilingual corpus – a list of 1:1 coupled sentences
▶ a phrase is a continuous sequence of tokens (for our purposes)
▶ extracts a list of (scored) equivalent phrases
▶ how phrases are extracted is not explained here

▶ also explores monolingual (target side) corpus to train language model
▶ simplified: lists of words with zero, one and two previous words
phrase based machine translation – simplified idea (2)

decoding/search for translation

▶ try to cover an input sentence with source-side of learned phrases
▶ target-side of selected phrases forms output sentence
▶ search is driven by phrase score and language model
▶ phrase model ensures translation correspondence
▶ language model tends to make output sentence grammatical

achieved abstraction

▶ phrases over sentences
phrase based machine translation – main issues

achieved abstraction

▶ phrases over sentences
▶ but no further generalization
▶ cannot even recognize an unseen form of a seen word in the language model

data sparseness

▶ in any available corpus we do not see all usages of all units (words)
▶ but we would like to see all translations in all their contexts in source language
▶ thus generalization is needed

Example
EBMT: close to mountains → close to X
generalization in language model

n-gram language model

- n-gram is n-tupple of tokens; e.g. n = 2
  \( w|_{\text{h}}: \text{řekla} \mid \emptyset \text{, řekla Že} \text{, půjde Že s} \text{půjde námi s} \text{námi} \)

- a sentence is scored on the basis of scores of n-grams it consists of (Bayes’ chain rule)

- usually n=3, 2 tokens of history, 1 predicted:
  \[ p(w_i|w_{i-2}w_{i-1}) \]

- higher n \( \rightarrow \) suffering more from data sparseness

- take into account also m-grams, \( 0 \leq m < n \) (smoothing)

smoothing with parts of speech

- if we have not seen the word in a given context of words, use at least the context of its POS

- \( p(\text{lesy}|\text{rozsáhlé}) = \cdots + \lambda_i p(\text{lesy}|\text{Adj.}) + \cdots \)
Carried out experiments’ basic facts

- employed data set: UMC 0.1 + extra set from ProjectSyndicate
- direction of translations: ru → cz, en → cz
- included methods: direct transfer, factored translation, both using Moses and related tools
- evaluation: Bleu, Gray-box evaluation
Data sources

Corus UMC 0.1

- Ufal Multilingual Corpus
- ProjectSyndicate articles new in 2009 extra 2,765 sentences tri-parallel

- numbers
  - LM sentences cz 92,233
  - TM sentences ru → cz 79,888
  - TM sentences en → cz 76,588
  - test set cz, en, ru 1,000
  - dev set cz, en, ru 750
Main steps

Data preparation

- factored TM training corpus
  - lemmatization and tagging
  - English&Russian by Tree-Tagger
  - Czech by J. Hajič tagger module in TectoMT
  - a lot of exercises with UNIX tools :-) 

Factored sentence snippets

prostě|prostě|Dg-----1A---- jsem|být|VB-S-1P-AA---
включая|включая|Sp-a президента|президент|Ncmsay
мбеки|мбеки|Vmip3s-a-p
the|the|DT visionaries|visionary|NNS would|would|MD
have|have|VH gotten|get|VVN nowhere|nowhere|RB
Main steps (2)

Running Moses

- direct transfer (simple)
- factored – two decoding paths
  1. (T) F.form $\rightarrow$ E.form, E.lemma, E.tag
  2. (T) F.lemma $\rightarrow$ E.lemma
     (T) F.tag $\rightarrow$ E.tag
     (G) E.lemma + E.tag $\rightarrow$ E.form
     + three separate LMs: for forms, lemmas and forms

Calling train-factored-phrase-model.perl

-lm 0:3:"$(WORK)/lm/cer.lctok.form.cz.blm"
-lm 1:3:"$(WORK)/lm/cer.lctok.lemma.cz.blm"
-lm 2:3:"$(WORK)/lm/cer.lctok.tag.cz.blm"
-translation-factors 0-0,1,2+1-1+2-2
-generation-factors 1,2-0
-decoding-steps t0:t1,t2,g0
explored settings

Russian  Czech

simple
form $\rightarrow$ form

factored1
(a) form $\rightarrow$ form, lemma, tag
(b) lemma $\rightarrow$ lemma
    tag $\rightarrow$ tag

factored2
(a) form $\rightarrow$ form
(b) lemma $\rightarrow$ lemma
    tag $\rightarrow$ tag
Evaluation of machine translation

evaluation criterion

► no single criterion
  ► preserves meaning
  ► outputs grammatical sentences
  ► what type of errors occur
  ► how much time/money does it take to correct the output, etc.
  ► we do not know user’s needs

our evaluation criterion

► automatic metric Bleu
► manual evaluation
  ► error analysis: missing word, extra word, bad word form, ...
  ► ranking – order translations of different systems
Evaluation – error analysis

- manual flagging of errors
- judge only of simple model (limited human resources)
- overview of errors

<table>
<thead>
<tr>
<th>Error Class</th>
<th>en→cs</th>
<th>ru→cs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disambiguation</td>
<td>9.3 %</td>
<td>8.8 %</td>
</tr>
<tr>
<td>Extra word</td>
<td>6.2 %</td>
<td>18.2 %</td>
</tr>
<tr>
<td>Word Form</td>
<td>49.0 %</td>
<td>22.0 %</td>
</tr>
<tr>
<td>Lexical Variant</td>
<td>5.4 %</td>
<td>5.7 %</td>
</tr>
<tr>
<td>Missed Auxiliary</td>
<td>0.8 %</td>
<td>1.9 %</td>
</tr>
<tr>
<td>Missed Content</td>
<td>6.6 %</td>
<td>20.1 %</td>
</tr>
<tr>
<td>Word Order Long</td>
<td>0.8 %</td>
<td>0.6 %</td>
</tr>
<tr>
<td>Word Order Short</td>
<td>4.6 %</td>
<td>0.6 %</td>
</tr>
<tr>
<td>Punctuation</td>
<td>13.9 %</td>
<td>2.5 %</td>
</tr>
<tr>
<td>Unknown</td>
<td>3.5 %</td>
<td>19.5 %</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>259 (100.0%)</td>
<td>159 (100.0%)</td>
</tr>
</tbody>
</table>
Evaluation – ranking

- which system produced the best translation?

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Type</th>
<th>factored1</th>
<th>factored2</th>
</tr>
</thead>
<tbody>
<tr>
<td>En→Cz</td>
<td>simple</td>
<td>2/8</td>
<td>9/6</td>
</tr>
<tr>
<td>Best/Second</td>
<td></td>
<td>4/6</td>
<td></td>
</tr>
<tr>
<td>ru→cz</td>
<td>simple</td>
<td>10/12</td>
<td>19/9</td>
</tr>
<tr>
<td>factorization</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- ru→cz, factored1 was the best the most times

- factorization helped particularly for translation from Russian
Evaluation – Bleu

- no significant improvement for English → Czech
- useful for Russian to Czech
- achieved Bleu scores in our experiments

### BLEU score on forms

<table>
<thead>
<tr>
<th>pair</th>
<th>simple</th>
<th>factored1</th>
<th>factored2</th>
</tr>
</thead>
<tbody>
<tr>
<td>en→cs</td>
<td>14.58±0.96</td>
<td>15.84±1.03</td>
<td>15.39±1.05</td>
</tr>
<tr>
<td>ru→cs</td>
<td>11.91±0.91</td>
<td>13.11±0.90</td>
<td>—</td>
</tr>
</tbody>
</table>

### BLEU score on lemmas

<table>
<thead>
<tr>
<th>pair</th>
<th>simple</th>
<th>factored1</th>
<th>factored2</th>
</tr>
</thead>
<tbody>
<tr>
<td>en→cs</td>
<td>24.16±1.10</td>
<td>24.77±1.18</td>
<td>24.99±1.16</td>
</tr>
<tr>
<td>ru→cs</td>
<td>15.98±0.97</td>
<td>18.06±0.92</td>
<td>—</td>
</tr>
</tbody>
</table>
Typical errors

Russian → Czech

- negation
  (cs ref) bez něhož nebylo možné sestavít
  (ru → cs): bez něhož bylo možné vytvořit

- reflexives
  (ru src) сумел уйти от
  (ru → cs) podařilo odejít od

English → Czech

- word order in possessive constructions
  (en src) mahmoud abbas ’s palestinian authority
  (cs ref) palestinskou samosprávou prezidenta mahmúda abbáse
  (en → cs) prezidenta mahmúda abbáse palestinské samosprávy
Both source languages → Czech

- Bad case after a preposition.
  (cs ref) podle indických vyšetřovatelů
  (en src) according to indian investigators
  (en → cs) podle indické řešitelů
  (ru src) согласно индийским экспертам
  (ru → cs) podle indickým experti
Conclusion

- less number of errors in errors flagging advices that translation from Russian is simpler
- it is also supported by manual ranking
- factorization is useful particularly for translation from Russian