Probes in a Taxonomy of Factored Phrase-Based Models

Ondřej Bojar, Bushra Jawaid, Amir Kamran surname@ufal.mff.cuni.cz

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

Pitfalls Highlighted

Notation and Taxonomy for Factored Models

Direct Translation

Baseline

- One search.
- All translation options come from one phrase table (t).

form (F) ○ → ○ (F) form tF-F

Single-Step

Generalized Morphology

- One search.
- Translation options constructed from several phrase (t) or generation (g) tables.

form (F) ○ → ○ (F) form
lemma (L) ○ → ○ (L) lemma tL-L+tT-T+gLaT-F
tag (T) ○ → ○ (T) tag

...with Alternative Decoding Paths

- The same input span covered by several paths (delimited by ":" in the notation).

Generalized and Can Fall Back

form (F) ○ → ○ (F) form
lemma (L) ○ → ○ (L) lemma
tag (T) ○ → ○ (T) tag

tL-L+tT-T+gLaT-F:tF-FaLaT

Two-Step

Avoids Combinatorial Explosion

- Two consecutive searches (delimited by "=").
- Single-best output from the first one passed to the second one.
- Each of the steps can use one or more factors.

form ○ → ○ LOF ○ → ○ form
○ MOT ○

tF-LOFaMOT = tLOFaMOT-F:tLOF-F

LOF ... lemma or form; MOT ... morphology or tag

Complex

...Anything You Can Imagine.

Direct Experiments

- With different target-side factors.
- With different language models.
...to improve target morphological coherence

Decoding Path	Language Models	BLEU
tF-FaLaT	form + lemma + tag	13.05±0.44
tF-FaT	form + tag	13.01±0.44
tF-FaLaT	form + tag	12.99±0.44
tF-F (baseline)	form	12.42±0.44
tF-FaT	form	12.19±0.44
tF-FaLaT	form	12.08±0.45

=> Additional factors bound to hurt unless used in a LM.

Vocabulary Sparseness

- The main problem when handling morphologically rich languages
- Direct setups can produce only forms seen in the parallel data

Spurious Ambiguity

Instead of just "police -> policie", we get more target-side entries due to the extra
police -> policie|sg
police -> policie|pl

Combinatorial Explosion of Translation Options

- Translation options are generated before the main search.
- All possible combinations of lemmas and tags are considered, best ones preserved.
- No context available during this generation.
=> Relevant translation options pruned before LM can promote them.

Two-Step Experiments

- With different properties captured in the "middle" language.
- With just one or two factors in the middle language.
...to avoid the combinatorial explosion while producing forms beyond what was seen in the parallel data.

- The division into two searches allows to consider linear context and apply a LM in the middle language.
- The first step performs lexical translation and reordering.
- The second search is monotone:
=> Can use max-phrase-length of just 1.
=> Can handle large number of possible target forms.
=> Can make use of larger (monolingual) data.

Input	Middle Language	Target
green	zelený+NSA-	zeleného (masc, genitive) zelenému (masc, dative) zelenou (fem, accusative)

Different Granularity of LOF and MOT

Word Form	LOF_0	LOF_1	MOT_0	MOT_1	MOT_2	Gloss
lidé	člověk	člověk	NNMP1-----A---1	NPA-	NMP1-A	people
by	být	by	Vc----------	c---	V-----	would
neočekávali	očekávat	očekávat	VpMP---XR-NA---	pPN-	VMP-RA	expect

One or Two Factors in the Middle Language

Middle Factors	1	2
	+	|
LOF_0 +/| MOT_0	11.11±0.48	12.42±0.48
LOF_1 +/| MOT_1	12.10±0.48	11.85±0.42
LOF_1 +/| MOT_2	11.87±0.51	12.47±0.51

- MOT1 performs better in single-factor setup.
- The more fine-grained MOTs 0 and 2 perform better when they can be disregarded.
- All our two-step setups below the baseline.

- Overgeneralization in Two-Step:
=> Omission of important source-side features in the middle language.

Single-Step Experiments

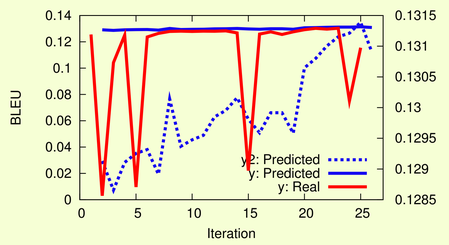
- Translate lemma and morphological tag separately.
- With different factors and LMs used.
- Optionally with an alternative decoding path: tF-FaLaT.

Decoding Paths	LMs	Avg. BLEU	Eff. Nbl. Size
tL-L+tT-T+gLaT-F:tF-FaLaT	F + L + T	13.31±0.06	12.24±1.33
tL-L+tT-T+gLaT-F	F + L + T	13.30±0.05	40.33±3.82
tL-L+tT-T+gLaT-F	F + T	13.17±0.01	39.91±2.58
tL-L+tT-T+gLaT-F:tF-FaLaT, 200-best-list	F + L + T	13.15±0.24	20.47±5.63
tF-FaLaT	F + L + T	13.13±0.06	34.28±3.08
tL-L+tT-T+gLaT-F:tF-FaLaT	L + T	13.09±0.06	16.65±1.07
tF-FaT	F + T	13.08±0.05	39.67±2.21
tL-L+tT-T+gLaT-F:tF-FaT	F + T	13.01±0.43	14.87±5.04
tF-F (baseline)	F	12.38±0.03	43.13±0.48
tL-L+tT-T+gLaT-F:tF-F	F	12.30±0.03	17.83±3.27

Effective N-Best List Size

- N-best lists often contain fewer than N unique items (e.g. different segmentations lead to the same output).
- Added ambiguity of target-side factors makes this worse.
- Alternative decoding paths make this far worse.

Predicted vs. Real Score in MERT



- MERT accummulates n-best lists from decoder runs at each iteration.
- Within MERT optimization, different weights are tested.
- These tests only predict from the n-best list what the final hypothesis could be.
- The actual search with such weights can lead to very different results, see the drops of the red curve.
- (Such drops are common for direct as well as complicated single-step setups.)
- MERT does recover from such drops in a few iterations.
- However, a drop too close to the iteration limit is unlikely to get recovered in time.
=> Weights obtained without convergence may be much worse than we expect.

Overgeneralization

- Unjustified independence assumptions reduce translation quality.
- Overgeneralization in Direct or Single-Step setups:

The|- police|sg investigate|pl
gets translated as: policie|sg vyšetřují|pl
instead of the correct: policie|sg vyšetřuje|sg

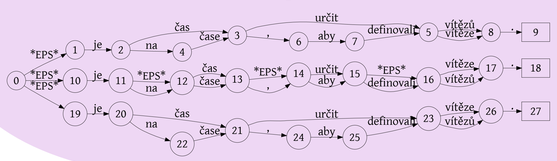
- Issue not noticed with very rich target-side morphology and small data.

TER Emphasizes Word Precision

- cu-poor-comb won in TER, but almost lost in BLEU and also manual evaluation.
- The reason are much shorter hypotheses of cu-poor-comb, essentially just the intersection of words produced by the combined systems.

System Combination

- Our implementation of system combination.
- Three Moses setups first produce primary hypotheses.
  - Alignment back to the source emitted, used to align hypotheses to each other.
- We combine them to a lattice similar to the ROVER setup.
  - Unlike in ROVER, our alignment can lead to multiword phrases.
- A fourth Moses selects the best path in the lattice.



Arc scores specify e.g.:
- the percentage of systems traversing that arc,
- indicators for each system, if it served as the skeleton
- indicator if the arc was traversed by the skeleton system

- Bad result due to too short outputs (even content words dropped).
...and this was due to a (1-TER)-optimized hypothesis in the combination.

Data

Dataset	Sents (cs/en)	Toks (cs/en)	Source
Small	197k parallel	4.2M/4.8M	CzEng 1.0 news
Large	14.8M parallel	205M/236M	CzEng 1.0 all
Mono	18M/50M	317M/1.265G	WMT12 mono

- All experiments used Small data only.
- All submitted systems use Large and Mono data.

Omitted Training Data

- We forgot to include WMT12's Czech-English parallel data (News Commentary and Europarl)
- This makes our system more comparable to CU-TectoMT.

Systems Submitted

		Our Scoring				matrix.statmt.org	
	Test Set	newstest-2011		newstest-2012			
	Metric	BLEU	TER*100	BLEU	TER*100	BLEU	TER
→cs	CU-POOR-COMB	–used–for–	–tuning–	14.17±0.53	64.07±0.53	14.0	0.741
	CU-BOJAR (tFaT-FaT, lex. r.)	18.10±0.55	62.84±0.71	16.07±0.55	65.52±0.59	15.9	0.759
	As ↑but towards 1-TER	16.10±0.54	61.64±0.59	14.13±0.54	64.28±0.55	–	–
	Large Two-Step	17.34±0.57	63.47±0.66	15.37±0.54	65.85±0.57	–	–
	Unused (tFaT-FaT, dist. reord.)	18.07±0.56	62.74±0.70	15.92±0.57	65.50±0.60	–	–
	Unused (tF-FaT, dist. reord.)	17.85±0.58	63.13±0.68	15.73±0.55	65.85±0.58	–	–
	Unused (tF-F, lex. reord.)	17.73±0.58	63.04±0.68	15.61±0.57	65.76±0.58	–	–
	Unused (tFaT-F, dist. reord.)	17.62±0.56	62.97±0.70	15.33±0.58	65.70±0.59	–	–
	Unused (tF-F, dist. reord.)	17.51±0.57	63.32±0.69	15.48±0.56	65.79±0.58	–	–
→en	CU-BOJAR (tF-F:tL-F, dist. reord.)	24.65±0.60	58.54±0.66	23.09±0.59	61.24±0.68	21.5	0.726
	Unused (tF-F, dist. reord.)	24.62±0.59	58.66±0.66	22.90±0.56	61.63±0.67	–	–

This work was supported by the project EuroMatrixPlus (FP7-ICT-2007-3-231720 of the EU and 7E09003+7E11051 of the Czech Republic) and the Czech Science Foundation grants P406/11/1499 and P406/10/P259. Presented at WMT12, Montreal.