LIMSI @ WMT'15 : Translation Task

Benjamin Marie[1,2,3], *Alexandre Allauzen*[1,2], *Franck Burlot*[1], *Quoc-Khanh Do*[1,2], *Julia Ive*[1,2,4], *Elena Knyazeva*[1,2], *Matthieu Labeau*[1,2], *Thomas Lavergne*[1,2], *Kevin Löser*[1,2], *Nicolas Pécheux*[1,2], *François Yvon*[1]

(1) LIMSI-CNRS, (2) UniversitÃľ Paris-Sud, (3) Lingua et Machina, (4) Centre Cochrane français

Systems Overview

SMT systems

- N-gram based: Ncode
- Phrase-based: Moses

Tools

- Word alignments: `fast_align`
- Language model: `lmplz`, 4-gram LMs pruning all singletons

SOUL

- Neural network language model and translation models with a Structured OUtput Layer used to rerank the n-best hypotheses produced by the decoder

Data

Corpus	Fr-En Sentences	Fr-En Tokens (Fr-En)	Ru-En Sentences	Ru-En Tokens (Ru-En)	Fi→En Sentences	Fi→En Tokens (Fi-En)
parallel data	24.3M	712.8M-597.7M	2.3M	45.7M-47.3M	2M	37.3M-51.7M
monolingual data		2.2B-2.7B		834.7M-2.7B		-2.7B

French-English

A new task

- Translate user-generated News discusssions

Internal development and test sets

- 3-fold cross-validation: split in 2 parts the 1,500 sentences of the official development set
- Each random split respects document boundaries:
  - Development set: ∼1,000 sentences
  - Test set: ∼500 sentences
  - Same proportion of documents from each source (*Le Monde* and *The Guardian*) in both sets

Domain adaptation

- No in-domain bilingual data, only in-domain monolingual data
- For translation table adaptation: subsampled the noisy Common Crawl and Giga Fr-En corpus, around 90% of all our bilingual data, using the Modified Moore-Lewis (Axelrod et al., 2011) filtering method (MML)
- For LM adaptation: log-linear combination of our large LM with a smaller one trained only on the monolingual in-domain corpus

N-best list reranking

- Rerank the N-best hypotheses of the decoder with features not used during decoding
- Features: IBM1, in-domain 6-gram POS LM, SOUL models, ratio of POS tag, word posterior probability

Results

MML effect

Configuration		Fr-En
baseline		29.33
before	10%	28.63
	25%	29.09
	50%	28.96
after	10%	29.14
	25%	29.31
	50%	29.11

LM adaptation

Configuration	Fr-En	En-Fr
w/o additional LM	29.15	29.56
w/ additional LM	29.33	30.22

Additionnal features

Feature sets	Fr-En	En-Fr
baseline	29.33	30.22
+ **IBM1**	29.24	30.25
+ **POSLM**	29.45	30.28
+ **SOUL**	30.20	31.15
+ **TagRatio**	29.33	30.30
+ **WPP**	29.40	30.20
`all`	30.45	31.25

Moses and Ncode results

System	in-house test Fr-En	in-house test En-Fr	official test Fr-En	official test En-Fr
Moses	29.33	30.22	32.16	**35.74**
Ncode	28.66	30.17	**32.85**	35.00

Russian-English

Preprocessing Russian

- Russian normalization: replace all case marks by the corresponding nominative inflection for nouns, pronouns and adjectives.

Postprocessing russian output

- Translate from English to Normalized Russian
- Retrieve the morphology of normalized words with a cascade of Conditional Random fields predicting:
  - POS-tags
  - Gender and number
  - Case
- Generate correct word form according to the former predictions.

Results

Russian-English

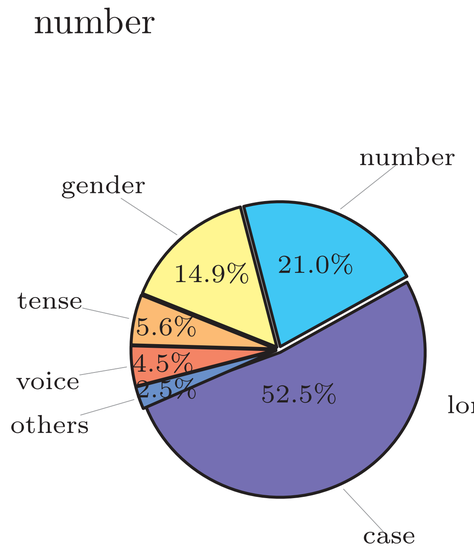
System	Moses	Ncode
Baseline	26.85	26.02
+ Normalized Ru	27.27	26.44
+ SOUL		27.28

English-Russian

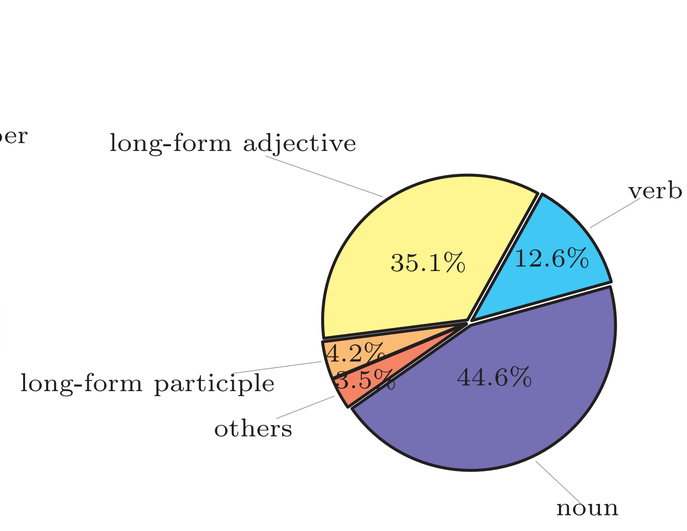
System	Moses	Ncode
Baseline	22.91	22.97
+ SOUL		24.08
En-Norm.Ru	26.35	26.12
En-Norm.Ru-Ru	19.99	19.88

Error analysis

- Errors made by the Ncode baseline system
- To identify errors at word-level Ncode output is aligned with the reference using METEOR
- Most of the morphological errors are related to case prediction and number



Incorrectly predicted inflections



Word form errors *wrt* POS

Finnish to English

- Preliminary experiments on morphological segmentation (with `Morfessor`) did not yield significative improvement.

Configuration	dev.	test
Baseline	13.2	12.8
+ large LM	16.1	15.7
+ Morph. segmentation	16.2	15.9

Conclusion

- For Fr-En, filtering the bilingual data did not bring any gains, while adding an in-domain language model yielded slight improvements
- For Ru-En, small improvements with a tailored normalization of Russian when translating into English