АБУМАТРАН
ABUMATRAN

MT for Finnish in the Abu-MaTran Project

Antonio Toral

FINMT 2016
12th September 2016
University of Helsinki

DCU prompsit Universitat d'Alacant Universidad de Alicante IEA ILSP

In This Talk

- Aims & Motivation
- What was done
- Lessons learnt
- Suggested ways forward

WMT 2015

Motivation

- Lack of resources
  - Web Crawling
    - obtain as much as possible relevant monolingual and parallel data
- Complex morphology
  - Morph segmentation

Web Crawling

1. Monolingual crawls (FI and EN) from .fi TLD
   - 1.7B FI and 2.0B EN words (40 days)

2. Parallel crawls on web domains with pages in both FI and EN with Bitextor and ILSP-FC
   - 2.8M segment pairs

How to combine datasets?

- Additional data (crawled and open subs) is noisy
  - Rank sentence pairs wrt devset (cross-entropy)
  - Divide in 3 sets
    - Pseudo in-domain (subset with lowest PPL)
      - FiEnWaC: top ¼ (404 EN, 3600 FI)
      - Opensubs: top 1/16 (702 EN, 7032 FI)
    - Remaining split in 2 equal parts: out-top, out-bottom

How to combine datasets?

- Additional data (crawled and open subs) is noisy
  - Ra… …ntropy)
  - Div…
    - …
    - … bottom

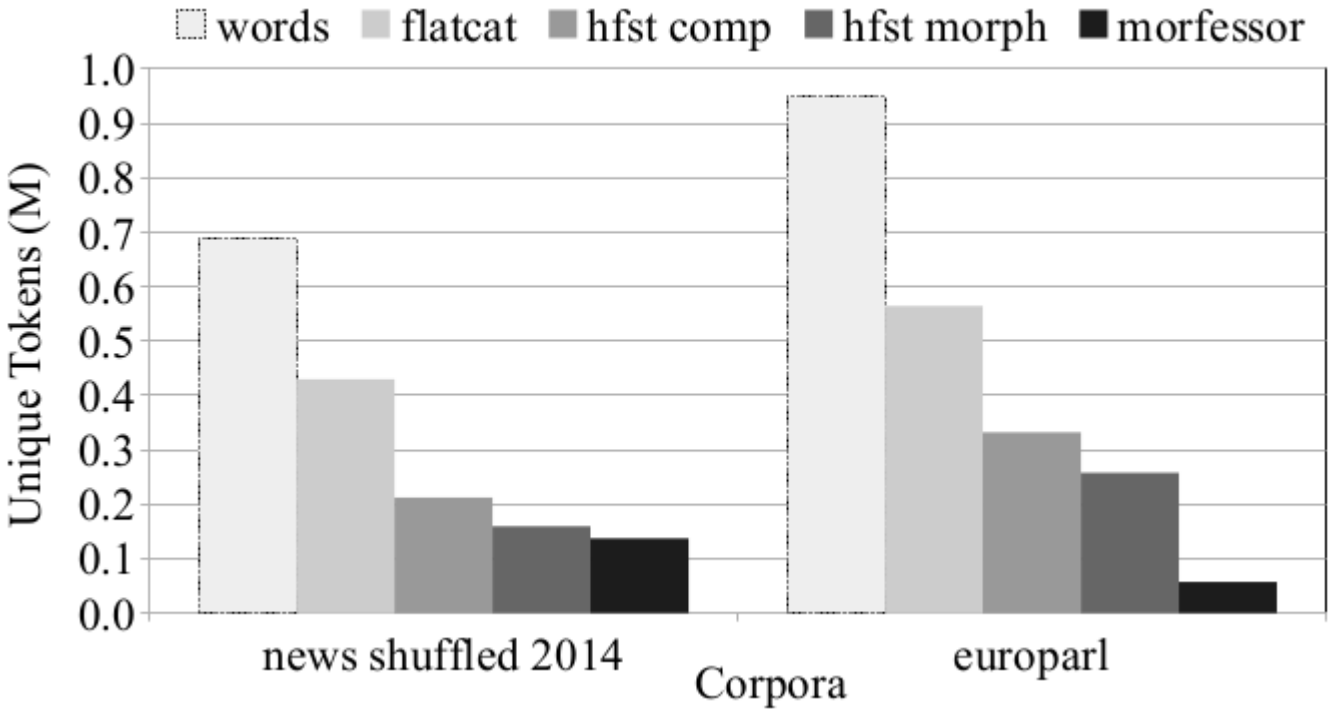
Corpus	Sentences (k)	Words (M) Finnish	English
	Constrained System		
Europarl v8	1,901.1	36.5	50.9
	Unconstrained System		
fienwac.in	640.1	9.2	13.6
fienwac.outt	838.9	12.5	18.1
fienwac.outb	838.9	13.9	18.1
osubs.in	492.2	3.6	5.6
osubs.outt	1,169.6	8.8	14.4
osubs.outb	1,169.6	7.8	13.0

Morph Segmentation

Segmenter	text
None	kuntaliitoksen selvittämisessä
hfst-comp	‘kunta→←liitoksen selvittämisessä
hfst-morph	kunta→←liitokse→←n selvittämise→←ssä
Flatcat	kun→←tali→←itoksen selvittämis→←essä
Morfessor	kun→←ta→←liito→←ksen selvittä→←misessä
Gloss	municipality+annexation.Gen examination.Ine
Translation	examination regarding municipal annexation

Morph Segmentation

- Pro: vocabulary reduction



Evaluation (auto)

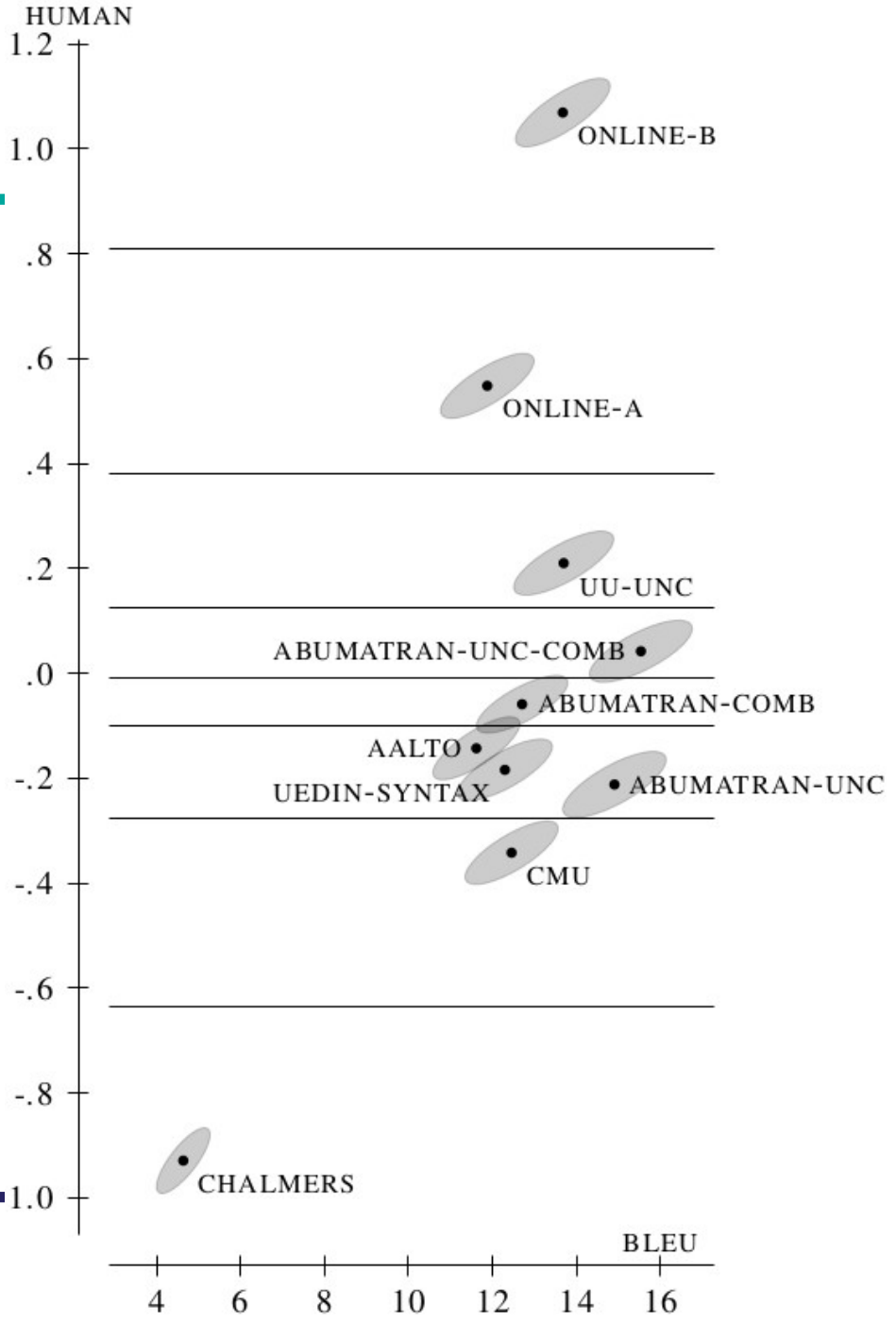
System	Dev BLEU	Dev TER	Test BLEU	Test TER
Phrase-Based	13.51	0.827	12.33	0.843
Factored Model	13.08	0.827	11.89	0.847
Hierarchical	13.05	0.822	12.11	0.830
HFST Comp	**13.57**	**0.814**	12.66	0.828
HFST Morph	13.19	0.818	**12.77**	**0.819**
Morfessor	12.21	0.860	11.58	0.864
Flatcat	12.67	0.844	12.05	0.849
Combination	14.61	0.786	13.54	0.801

Evaluation (auto)

System	Dev BLEU	Dev TER	Test BLEU	Test TER
Phrase-Based	13.51	0.827	12.33	0.843
Factored Model	13.08	0.827	11.89	0.847
Hierarchical	13.05	0.822	12.11	0.830
HFST Comp	**13.57**	**0.814**	12.66	0.828
HFST Morph	13.19	0.818	**12.77**	**0.819**
Morfessor	12.21	0.860	11.58	0.864
Flatcat	12.67	0.844	12.05	0.849
Combination	14.61	0.786	13.54	0.801

System	Dev BLEU	Dev TER	Test BLEU	Test TER
Phrase-Based	**16.16**	0.804	**16.07**	0.801
HFST Comp	15.80	**0.796**	15.06	**0.800**
Combination	17.25	0.776	16.38	0.779

Evaluation (human)



Evaluation (human)

Post-mortem experiment

#	Score	Range	System
1	0.529	1-2	combo-all
2	0.414	1-2	combo-rb
3	-0.943	3-3	combo-unsup

Tommi A Pirinen, Antonio Toral and Raphael Rubino. **Rule-Based and Statistical Morph Segments in English-to-Finnish SMT**. 2nd International Workshop on Computational Linguistics for Uralic Languages. 2016.

WMT 2016

Auto != Human Results at WMT15

- Correlation of metrics with human judgements for Finnish
  - BLEU
  - chrF
- Tune on chrF (combo, ~~Moses~~)
- Look at a range of metrics during development
  - BLEU, chrF, TER, METEOR

Trade-off segm. – non segm. (SMT)

System	BLEU	TER	chrF1
No segmentation	0.1444	0.7775	49.63
`Omorfi`	0.1501 ↑	0.7717 ↑	**51.13** ↑
`Omorfi` + BPE	**0.1536** ↑	**0.7679** ↑	50.99 ↑

NMT

- Train
  - EU (1.9M sentence pairs)
  - Backtranslated Finnish news (6.7M)
- Segmentation
  - None vs unsupervised (BPE) vs rule-based
- Treatment of UNKs to preserve NEs
- Ensemble

NMT and SMT combo

- Despite gap in performance (NMT>>SMT)
  - can SMT still be useful on top of NMT?

NMT and SMT combo

- Despite gap in performance (NMT>>SMT)
  - can SMT still be useful on top of NMT?

System	BLEU	TER	chrF1
Best SMT	0.1562	0.7644	51.04
Best NMT	0.1830	0.7411	52.43
Combo (BLEU)	0.1638	0.7298 ↑	51.75
Combo (chrF1)	0.1767	**0.7241** ↑	52.37
Reranked (BLEU)	0.1791	0.7257 ↑	52.38
Reranked (chrF1)	**0.1845**	0.7290 ↑	**52.65** ↑

Ideas Going Forward

Some Ideas

- Keep what works
  - Rule-based segmentation
  - NMT
- How to make the best of the (parallel) data?
  - EU, backtranslated news, FiEnWaC, OPUS
- Better metrics
  - Finnish in TERp / METEOR
- Analysis
  - On what are we improving?
  - What remains weak (or even has degraded)?

АБУМАТРАН
ABUMATRAN

Thanks!

MT for Finnish in the Abu-MaTran Project

Antonio Toral

DCU prompsit Universitat d'Alacant Universidad de Alicante IEA ILSP