Experiments on domain-specific Statistical Machine Translation at the European Parliament

ASLIB 2012

Alexandros Poulis
Project Manager
Intrasoft/European Parliament
alexandros.poulis@ext.europarl.europa.eu

Konstantinos Chatzitheodorou
PhD Candidate, Aristotle University of Thessaloniki
kchatzitheodorou@gmail.com

OVERVIEW

- Why MT @ EP?
- Experimental Setup
- MT systems compared
- Evaluation
- System combination
- Conclusions and future work

Why MT @ EP?

- **Because we need it:**
  - Increasing need for translation
  - Q1 of 2010: 43,963 source pages to be translated
  - Q1 of 2012: 60,275 source pages to be translated
- **Because we can:**
  - Availability of in-house corpora
  - Most translations are stored in translation memories which can be used as corpora for MT
- **Fact: 23 official languages all equally important**
  - Every member has the right to speak in the official language of her/his choice
  - Transparency and accessibility for EU citizens
- **Fact: 506 possible language combinations**

EXPERIMENTAL SETUP

- **Domain of experimentation:**
  - Verbatim reports of EP proceedings (CRE)
- **Language pair: EN-EL**
- **Objective:**
  - Improvement of the MT output, combining the output of MT systems trained with different kind of corpora

EXPERIMENTAL SETUP: System 1

- **Training corpus: Europarl V6 (P. Koehn)**
  - in-domain data
- **Tuning corpus: 1.872 CRE sentences**
- **Phrase based open-source Moses toolkit**
- **GIZA++ for the word alignment training**
- **SRILM for the 7-gram language models**

	Corpus	Sentences	Words		Distinct words	
			EN	EL	EN	EL
training	Europarl	1.064.544	27.357.281	27.359.635	119.817	248.482
tuning	CRE	1.872	43.834	45.035	4.930	8.320

EXPERIMENTAL SETUP: System 2

- **Training corpus: EURAMIS translation memories of European Parliament and European Commission**
  - EP but no CRE data
- **Tuning corpus: 1.977 EURAMIS sentences (not in the training corpus)**
- **Phrase based open-source Moses toolkit**
- **GIZA++ for the word alignment training**
- **SRILM for the 7-gram language models**

	Corpus	Sentences	Words		Distinct words	
			EN	EL	EN	EL
training	EURAMIS	8.643.223	159.026.130	166.813.972	706.234	1.028.434
tuning	EURAMIS	1.997	55.466	58.557		

EXPERIMENTAL SETUP: System 3

- **Training corpus: EURAMIS translation memories of European Parliament and European Commission**
  - EP but no CRE data
- **Tuning corpus: 1.872 CRE sentences (in-domain)**
- **Phrase based open-source Moses toolkit**
- **GIZA++ for the word alignment training**
- **SRILM for the 7-gram language models**

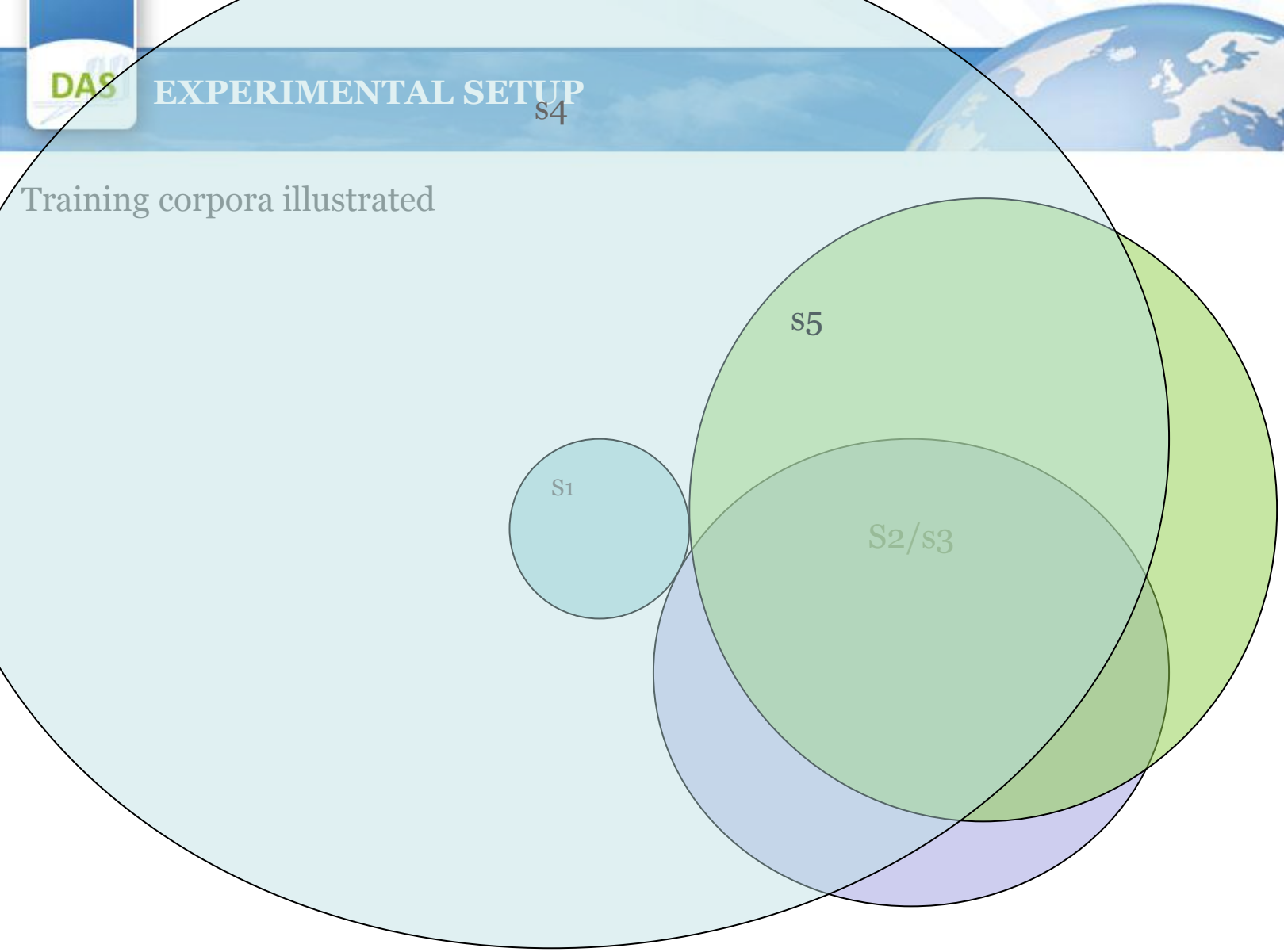
	Corpus	Sentences	Words		Distinct words	
			EN	EL	EN	EL
training	EURAMIS	8.643.223	159.026.130	166.813.972	706.234	1.028.434
tuning	CRE	1.872	43.834	45.035	4.930	8.320

EXPERIMENTAL SETUP: Systems 4 & 5

- **Free online MT system (S4)**
- **European Commission's MT system (S5)**
  - parallel corpus extracted by the translation memories and other bilingual recourses

EXPERIMENTAL SETUP

Training corpora illustrated



EVALUATION

- **Test corpus**
  - CRE content

Corpus	Sentences	Words		Distinct words	
		EN	EL	EN	EL
CRE	541	12.405	12.937	2.432	3.611

EVALUATION : AUTOMATIC

- **BLEU scores**
- **Test set 541 CRE sentences (12.405 EN & 15.937 EL words)**
- **One single reference translation per sentence**
- **In-domain tuning data yielded worse BLEU scores for two systems trained on the same corpora (S2>S3)**

MT System	BLEU score
S1	23.63
S2	19.42
S3	13.68
S4	33.74
S5	23.45

EVALUATION : HUMAN

- **Linguistic analysis by a Greek native speaker (linguist)**
- **Error Types in a set of 100 segments**

Error type	Occurrences S1	Occurrences S5
Word order		
- Single word	11	15
- Sequence of words	42	52
Incorrect word(s)		
- Wrong lexical choice	40	24
- Wrong terminology choice	10	8
- Incorrect form	38	44
- Extra word(s)	0	14
- Missing word(s)	50	10
- Style	10	0
- Idioms	2	2
Untranslated word(s)	4	2
Punctuation	5	10
Letter case	2	1
Other	1	1

SYSTEM COMBINATION

- **MT system combination**
  - Multi-Engine MT software (MEMT) (Heafield and Lavie, 2010)
- **Parameter weights**
  - Tuning corpus: 500 segments of CRE documents
  - 7-gram language model of Europarl corpus
- **MT outputs selected**
  - S1 & S5 (two systems with the higher BLEU score)
- **Evaluation**
  - BLEU scores of the same test corpus
- **Result**
  - The combination of the two systems provided an additional increase of 0.2 BLEU points (S1 23.63, S5 23.45, MEMT 23.83)

CONCLUSIONS

- Availability of in-domain training data improved BLEU scores even in a domain with not low amounts of repetitive text
- In-domain tuning data yielded worse BLEU scores for two systems trained on the same corpora (S2>S3)
- System combination helped us improve the BLEU scores compared to the best performing system
- The in-domain system (s1) produced better word-order output while the general-domain s5 with much more data made significantly better lexical choices and had a much greater coverage than s1 according to the human evaluation.

FUTURE WORK

- Run a large-scale human evaluation campaign to estimate the benefits of MT and define use-cases
- Combine Euramis data with Europarl corpus (in-domain)
- Create in-house corpora from available document resources and enhance the available MT data. Most corpora will be provided to the research community.
- Run experiments in other domains

Thank you!