LIMSI @ WMT'14 Medical Translation Task

Nicolas Pécheux[1,2], *Li Gong*[1,2], *Quoc Khanh Do*[1,2], *Benjamin Marie*[2,3], *Yulia Ivanishcheva*[2,4], *Alexandre Allauzen*[1,2], *Thomas Lavergne*[1,2], *Jan Niehues*[2], *Aurélien Max*[1,2], *François Yvon*[2]

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

Highlights

- Subtask of **sentence translation from summaries**, English → French

```
In what circumstances do granulomatous and eosinophilic gastritis occur?
What are the etiologies of dysphagia in gastroesophageal reflux disease?
```

- Successful approach that makes use of **two flexible translation systems**

Data sources

Corpus	Tokens (en)	weight
Coppa	10M	-3
Emea	6M	26
Pattr-Abstracts	20M	22
Pattr-Claims	32M	6
Pattr-Titles	3M	4
UMLS	8M	-7
Wikipedia	17k	-5
NewsCommentary	4M	6
Europarl	54M	-7
Giga	260M	27
all	397M	33

- **Combining both data sources** drastically boosts performance

	Devel	Test
medical	42.2± 0.1	39.6± 0.1
WMT'13	43.0± 0.1	41.0± 0.0
both	48.3± 0.1	45.4± 0.0

BLEU scores obtained by Ncode

Part-of-Speech Tagging

- Medical data exhibit **different syntactic constructions and a specific vocabulary**
- We use a specific model trained on medical data

PoS tagging	Devel	Test
Standard	47.9± 0.0	44.8± 0.1
Specialized	48.3± 0.1	45.4± 0.0

Proxy Test Set

- Only a small development set is available (500 sentences)
- This makes both **system design and tuning challenging**
- We created an internal dev/test set (LmTest) by extracting sentences from Pattr-Abstracts

Devel	LmTest	NewsTest12	Test
48.3± 0.1	46.8± 0.1	26.2± 0.1	45.4± 0.0
41.8± 0.2	48.9± 0.1	18.5± 0.1	40.1± 0.1
39.8± 0.1	37.4± 0.2	29.0± 0.1	39.0± 0.3

Error Analysis

	extra	*missing*		*incorrect*					*unknown*		
	word	content	filler	disamb.	form	style	term	order	word	term	**all**
SysComb	4	13	20	47	62	8	18	21	1	11	**205**
OTF+VSM+Soul	4	4	31	44	82	6	20	42	3	12	**248**

Manual error analysis following (Vilar et al., 2006) for the first 100 test sentences.

Systems

MIRA

Ncode — bilingual *n*-gram approach to SMT

OTF — on-the-fly estimation of the parameters of a standart phrase-based model

VSM — Vector space model to perform domain adaptation

Soul — Continous space models working on top of conventional language models (reranking); adapted language model (LM*)

SysComb — Combination of both systems (reranking)

	Devel	Test
Ncode	48.5	45.2
+ Soul LM*	49.8	45.9
+ Soul LM*+ TM	50.1	47.0
OTF	46.6	42.5
+ VSM	46.9	42.8
+ Soul LM*	48.4	44.2
+ Soul LM*+ TM	49.7	44.9
SysComb	50.7	46.5

- Ncode outperforms OTF by 2.8 BLEU points
- Vector space model does not yield here any improvement
- **Continous space language models yield gains** of up to 2 BLEU points
- System combination **gain does not transfer to the test set**

Conclusions

- **Moderate to high-quality** translations
- **Lack of an internal test** challenging
- **More careful integration of medical terminology** proved necessary