Machine Translation of Medical Texts in the Khresmoi Project

Ondřej Dušek, Jan Hajič, Jaroslava Hlaváčová, Michal Novák, Pavel Pecina, Rudolf Rosa, Aleš Tamchyna, Zdeňka Urešová, Daniel Zeman

{odusek,hajic,hlavacova,mnovak,pecina,rosa,tamchyna,uresova,zeman}@ufal.mff.cuni.cz
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics
Charles University in Prague

Introduction

- the participation of the Charles University team in the WMT 2014 Medical Translation Task
- developed within the Khresmoi project
- our **primary goal** was to **set up a baseline for both subtasks**: summary and query translation

Data selection

- train an **in-domain LM$_M$** (medical) and a **general-domain LM$_N$** (news)
- for each sentence *s*, compute **per-word cross-entropies** given the language models
- select *s* as **(pseudo)-in-domain if H$_M$(*s*) < H$_N$(*s*)**
- for parallel data, use only the English side

System design

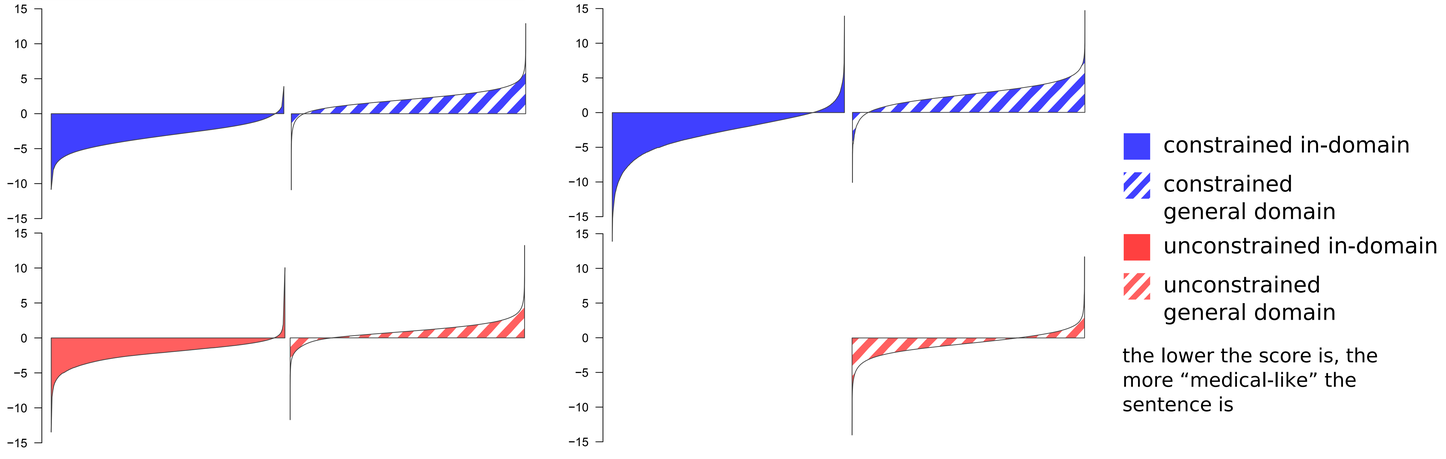
- **based on** the phrase-based **Moses** system
- combining "medical-like" data from the medical and general domain
- CUNI primary: **linear interpolation**
- CUNI contrastive: **concatenation**

Monolingual data

source	English	Czech	German	French
UMLS	7,991	63	24	37
Wiki	26,945	1,784	10,232	8,376
PatTR	121,592		53,242	54,608
AACT	13,341			
DrugBank	953			
GENIA	557			
GREC	62			
FMA	884			
PIL	662			
News	2,194,609	608,243	1,668,053	797,961
News-Europarl	68,375	19,250	60,012	75,925
Gigaword	3,869,122			963,571
CLEF	424,109			
HON	2,806,566	3,167	347,905	892,091
non-HON	44,596	33,180	13,975	16,820
PatTR	618,084		339,595	204,025

Parallel data

	Czech–English			German–English			French–English		
dataset	pairs	source	target	pairs	source	target	pairs	source	target
EMEA	1,053	13,872	14,378	1,108	13,946	14,953	1,092	17,605	14,786
UMLS	1,441	4,248	5,579	2,001	6,613	8,153	2,171	8,505	8,524
Wiki	3	5	6	10	19	22	8	19	17
MuchMore				29	688	740			
PatTr				1,848	102,418	106,727	2,201	127,098	108,665
COPPA							664	49,016	39,933
CommonCrawl	161	3,603	4,059	2,399	56,516	60,483	3,244	95,253	83,302
News+Europarl	793	18,270	21,269	2,121	56,301	58,920	2,190	71,130	61,191
CzEng	14,833	204,837	235,177						
UN							12,886	449,279	372,627
GigaFrEn							22,520	854,353	694,394
PatTr				9,320	525,782	574,373	10,967	703,165	595,376
COPPA							2,841	258,826	212,846



Results

- query translation, normalized, lowercased, 1-TER

	Czech → English		English → Czech	
CUNI primary	62.53±2.84	65.02±2.99	51.34±2.51	52.89±2.40
CUNI contrastive	**62.56±2.99**	**65.52±2.26**	**51.99±2.40**	50.34±2.47
UEDIN primary	55.66±3.06		37.85±2.72	
UEDIN contrastive			24.96±3.50	
ONLINE		64.05±2.97		**58.10±2.50**
	German → English		English → German	
CUNI primary	**63.68±2.34**	**65.52±2.26**	**42.39±3.21**	46.12±2.74
CUNI contrastive	62.87±2.39	62.92±2.32	41.24±3.21	41.24±3.21
POSTECH primary	47.84±2.82		39.89±3.14	
POSTECH contrastive	53.33±2.55		34.92±3.40	
UEDIN	53.76±3.48		39.08±3.42	
ONLINE		60.86±3.22		**50.18±2.95**
	French → English		English → French	
CUNI primary	72.30±2.63	**74.49±2.45**	63.74±2.35	63.61±2.45
CUNI contrastive	71.25±2.76	73.08±2.57	61.43±2.37	**65.39±2.57**
DCU primary	**75.86±2.37**		**63.84±2.47**	
DCU contrastive	63.51±3.21			
UEDIN	54.31±3.17		48.52±4.07	
ONLINE		72.59±2.61		64.06±2.62

Conclusion

- successful **pseudo-in-domain data selection**
- linear model interpolation good for incorporating additional data

http://khresmoi.eu

Supported by the EU FP7 grant n°257528 (KHRESMOI), the Czech Sience Foundation grant n°P103/12/G084, and SVV project number 260 104. Resources are distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (LM2010013).

Presented at WMT, June 26th, 2014, Baltimore, USA