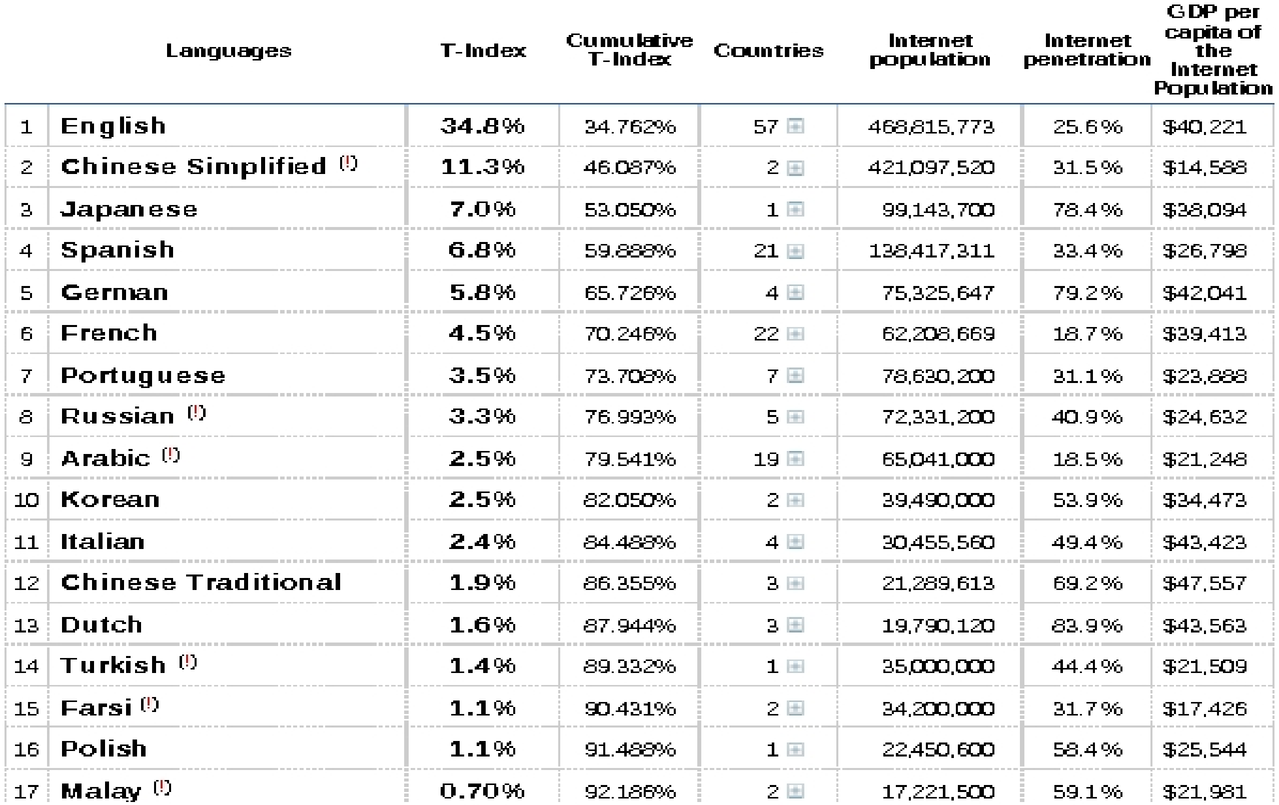
Bootstrapping Arabic-Italian SMT through Comparable Texts and Pivoting

Mauro Cettolo, Nicola Bertoldi & Marcello Federico
FBK, Trento - Italy

EAMT
May 30-31, 2011
Leuven, Belgium

Problem

Parallel data are scarce even for socially and economically relevant language pairs



	Languages	T-Index	Cumulative T-Index	Countries	Internet population	Internet penetration	GDP per capita of the Internet Population
1	**English**	**34.8%**	34.762%	57 ⊞	468,815,773	25.6%	$40,221
2	**Chinese Simplified** (!)	**11.3%**	46.087%	2 ⊞	421,097,520	31.5%	$14,588
3	**Japanese**	**7.0%**	53.050%	1 ⊞	99,143,700	78.4%	$38,094
4	**Spanish**	**6.8%**	59.888%	21 ⊞	138,417,311	33.4%	$26,798
5	**German**	**5.8%**	65.726%	4 ⊞	75,325,647	79.2%	$42,041
6	**French**	**4.5%**	70.246%	22 ⊞	62,208,669	18.7%	$39,413
7	**Portuguese**	**3.5%**	73.708%	7 ⊞	78,630,200	31.1%	$23,888
8	**Russian** (!)	**3.3%**	76.993%	5 ⊞	72,331,200	40.9%	$24,632
9	**Arabic** (!)	**2.5%**	79.541%	19 ⊞	65,041,000	18.5%	$21,248
10	**Korean**	**2.5%**	82.050%	2 ⊞	39,490,000	53.9%	$34,473
11	**Italian**	**2.4%**	84.488%	4 ⊞	30,455,560	49.4%	$43,423
12	**Chinese Traditional**	**1.9%**	86.355%	3 ⊞	21,289,613	69.2%	$47,557
13	**Dutch**	**1.6%**	87.944%	3 ⊞	19,790,120	83.9%	$43,563
14	**Turkish** (!)	**1.4%**	89.332%	1 ⊞	35,000,000	44.4%	$21,509
15	**Farsi** (!)	**1.1%**	90.431%	2 ⊞	34,200,000	31.7%	$17,426
16	**Polish**	**1.1%**	91.488%	1 ⊞	22,450,600	58.4%	$25,544
17	**Malay** (!)	**0.70%**	92.186%	2 ⊞	17,221,500	59.1%	$21,981

T-index: a combination of the Internet population and its estimated GDP per capita.

From www. translated.net/ /en/languages-that-matter

Experimental framework:

- under-resourced language pairs (Arabic–Italian)
- not ready-to-use training data (different nature, comparable texts, ...)

Research directions:

- automatic detection and extraction of parallel texts from the Web
- translation using pivot languages

Outline

- New *benchmark* developed by extending two Arabic→English NIST evaluation sets with Italian (and French) translations, from the source language by experts
- Many *direct SMT systems* have been developed:
  - from source to target language (Arabic→Italian)
  - from source to pivot languages (Arabic→English)
  - from a pivot to target (English→Italian)

  Methods for *exploitation* of *comparable texts* have been applied
- The *pivot* method known as *composition*, called *transfer* by Wu and Wang (2009), has been experimentally investigated

Benchmark

- a *professional translation* company was asked to translate the Arabic side into Italian (and French) of the sets provided for the 2009 MT NIST evaluation campaign - Arabic→English task
- one translation per sentence has been produced (i.e. *single reference*)
- the translation from Arabic *avoided* any *bias* towards English

Some statistics (word counts are given in thousands):

set	#sent.	Arabic		English		French		Italian	
		$\vert W\vert$	$\vert V\vert$	$\vert W\vert$	$\vert V\vert$	$\vert W\vert$	$\vert V\vert$	$\vert W\vert$	$\vert V\vert$
eval08-NW	813	21.9	7.8	29.1	4.9	33.2	4.9	32.0	5.7
eval09-NW	586	17.5	6.4	23.1	3.9	26.7	4.4	25.1	4.8

$|W|$ = text size $|V|$ = vocabulary size

Comparable corpora for SMT

General Scheme for collecting parallel data from comparable data:

1. cluster multilingual documents, by metadata, heuristics, IR ...
2. split documents into sentences
3. pair sentences across documents, by length, lexical overlap, word alignment ...
4. filter sentence or fragment pairs which align very well

Our approach fits this scheme and it is new on some aspects:

- Document pairing
- Mining parallel fragments

Document pairing

Problem: pair documents likely including parallel texts

Assumptions: documents include a title + baseline MT system

- Methods tested share the translation of titles from the language `A` into the `B`:
  - θ: documents paired if titles closer (e.g. wrt PER) than a threshold θ
  - NB: added a constrained translation for feeding a NB classifier
  - IR: indexed `B` documents are retrieved with translated `A` titles

Exps on 30K Italian/English docs from EuroNews:

method	%P	%R	$\%F_1$
θ	20.8	16.4	18.4
NB	26.8	25.3	26.0
IR	73.2	73.0	73.1

Mining parallel fragments

Novel method for collecting parallel fragments from comparable documents:

1. source document paired to each sentence of the target document
2. partial phrase-based alignment between the paired texts
3. aligned phrases iteratively merged into blocks on the basis of simple heuristics

→ final aligned blocks are the parallel fragments

Exps on ACL WMT 2010 German→English task (IWSLT 2010):

training data			%BLEU
baseline	additional		
running words	running words	type	
2.5M	-	-	17.6
	0.5M	fragments(EN)	18.5
	0.5M	sentences(EP)	17.9
	2.0M	sentences(EP)	18.3

Direct systems: training data

ArIt-fbk	$\lvert W \rvert$		trained
type	ar	it	models
web parallel sent.	1.4M	1.4M	
web parallel frag.	1.8M	1.6M	
total		3.0M	LM
total clean	3.0M	2.8M	TM RM
web monol. sent.		1.06G	LM

EnIt-fbk	$\lvert W \rvert$		trained
type	en	it	models
web parallel sent.	24.2M	24.1M	
web parallel frag.	2.7M	2.8M	
total		27.0M	LM
total clean	23.3M	23.5M	TM RM
ep5+acquis clean	70.0M	70.0M	TM RM
web monol. sent.		1.06G	LM

ArEn-fbk system developed on data provided for the NIST 2009 evaluation campaign

Direct systems: performance

Performance on the eval09-NW set of the direct systems developed for the translation from Arabic into Italian via pivoting (eval08-NW used for tuning):

system id	translation direction	training data #words (source)	%BLEU (4 refs)	%BLEU (1 ref)
ArIt-fbk	ar→it	3.0M	-	13.1
ArIt-ggleTrnslt		?	-	19.2
ArEn-fbkNist09	ar→en	147.2M	54.3	35.3
ArEn-ggleTrnslt		?	55.5	33.5
EnIt-fbk	en→it	93M	-	21.0
EnIt-ggleTrnslt		?	-	19.2

suffix `ggleTrnslt` = Google Translate - as it was in January 2011

Pivot systems: performance

Performance on the eval09-NW set of the pivot-based systems for the translation from Arabic into Italian:

translation direction	paired systems			%BLEU
ar→it	ArEn-fbkNist09	⊗	EnIt-fbk	19.5
	ArEn-ggleTrnslt	⊗	EnIt-ggleTrnslt	18.2

- our pivoting is effective (19.5 vs. 19.2 by EnIt-ggleTrnslt)
- ggleTrnslt: 19.2 by “direct” vs. 18.2 by “single” composition

  ⇒ this suggests us to further work on our pivot chain from Arabic to Italian, e.g. by including more pivot languages (French) and by combining multiple systems

Work in progress

- Daily crawling of data from news web sites
- Efficiency in fragment extraction
- Improving direct SMT systems
- Adding French as pivot language for Arabic→Italian
- Synthetic and triangulation (open issue: reordering model) pivot translation