On the Use of Words Alignments to Enhance Bitext Compression

Miguel A. Martínez-Prieto, Joaquín Adiego, Felipe Sánchez-Martínez, Pablo de la Fuente and Rafael C. Carrasco

Universidad de Valladolid

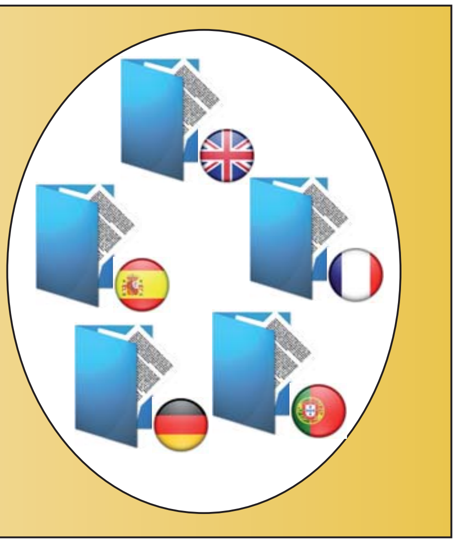
Universitat d'Alacant
Universidad de Alicante

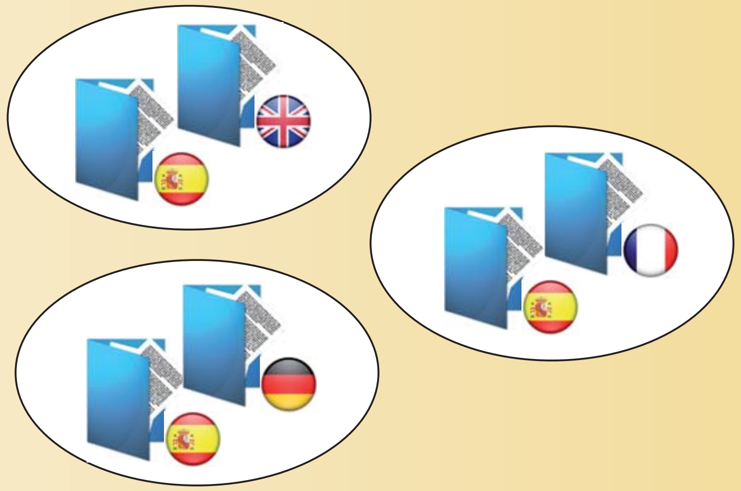
19th Data Compression Conference. Snowbird, Utah, USA. March 16-18, 2009

Multilingual Parallel Corpora

Represents the same information in different languages, i.e.:

- Official texts of the European Union.
- Daily newspapers in Spain.
- Information on Internet.





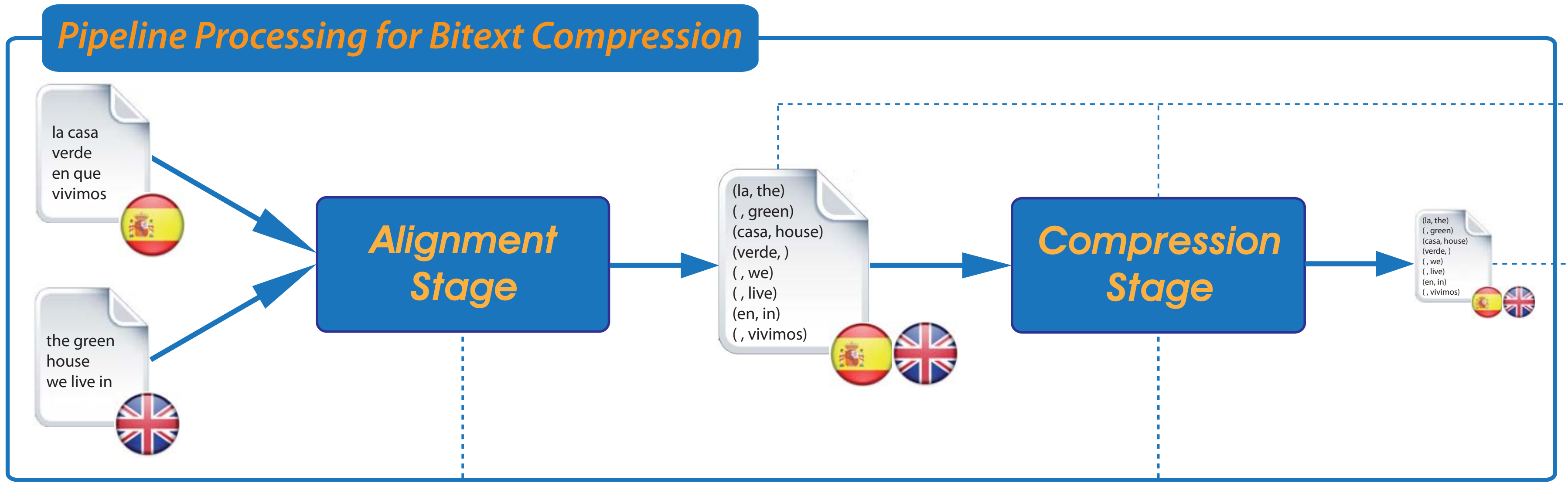
Bitexts: Bilingual Parallel Corpora

Two texts that are mutual translations.

Bitexts Compression

Bitexts show two different representations of the same information.
A shared representation of both texts should improve compression.

Pipeline Processing for Bitext Compression



Alignment Stage

Obtains a biword representation of the bitexts.

A **biword** is a pair made of two words, each one from a different text or empty, that are mutual translations in the bitext.

We use **GIZA++** to compute the word alignments from which biwords are generated.

Compression Stage

Mapping biwords on a limited-size dictionary: **mppm** variation. Features:

- Each biword is encoded using a 2-bytes code.
- When dictionary is full, a LRU policy is applied in order to replace biwords.
- Each 2-bytes code is compressed using PPMDi.
- Boosting on PPM.

Dictionary & output sizes evolution

	size in MB	Dictionary Size (Different Symbols)				Output Size (Total Symbols)			
		S	T	(S,T)	$\frac{S+T}{(S,T)}$	S	T	(S,T)	$\frac{S+T}{(S,T)}$
es-en	1	9 168	6 695	21 301	**0.745**	92 912	91 105	133 260	**1.381**
	5	21 840	14 365	61 102	**0.593**	472 112	458 367	676 992	**1.374**
	10	30 485	19 464	93 544	**0.534**	943 408	915 649	1 353 562	**1.373**
	20	42 212	26 240	142 888	**0.479**	1 899 559	1 847 753	2 727 579	**1.374**
	40	56 804	35 101	213 373	**0.431**	3 785 111	3 682 455	5 425 094	**1.376**
	60	67 437	41 710	269 299	**0.405**	5 708 010	5 557 560	8 162 767	**1.380**
	100	81 866	51 351	347 866	**0.383**	9 584 394	9 320 814	13 447 743	**1.406**
es-ca	1	15 593	14 938	19 336	**1.579**	106 264	114 691	125 778	**1.757**
	5	38 365	36 895	52 539	**1.432**	520 810	561 398	616 143	**1.756**
	10	54 114	52 256	78 825	**1.349**	1 039 047	1 118 617	1 229 030	**1.756**
	20	75 885	73 632	116 594	**1.282**	2 063 032	2 221 166	2 439 881	**1.756**
	40	107 245	104 852	175 990	**1.205**	4 053 672	4 359 401	4 796 340	**1.754**
	60	129 067	126 233	220 892	**1.156**	5 960 013	6 402 697	7 053 345	**1.753**
	100	161 131	159 216	292 994	**1.093**	9 586 737	10 282 568	11 359 915	**1.749**

Comparison and compression ratios achieved with different bitext collections

	MB	gzip	bzip2	PPM	mPPM	Bi-mPPM	
						NoPack	Pack
es-gl	1	20.97%	13.30%	11.46%	11.04%	8.01%	7.19%
	5	20.80%	13.09%	11.43%	10.28%	6.32%	5.81%
	10	20.84%	13.08%	11.33%	10.11%	5.94%	5.53%
es-ca	1	37.64%	28.93%	26.65%	26.51%	18.81%	17.96%
	5	37.41%	27.92%	25.90%	24.56%	15.86%	15.35%
	10	37.31%	27.70%	25.75%	23.99%	14.95%	14.58%
	20	37.28%	27.65%	25.75%	23.95%	14.51%	14.20%
	40	37.41%	27.74%	25.85%	23.93%	14.43%	14.13%
	60	37.28%	27.56%	25.71%	23.69%	14.28%	13.99%
	100	37.09%	27.21%	25.43%	23.30%	14.05%	13.78%
es-en	1	31.75%	24.28%	20.65%	21.39%	21.53%	21.44%
	5	31.61%	22.25%	20.50%	19.29%	19.50%	19.47%
	10	31.50%	22.22%	20.36%	18.95%	18.94%	18.91%
	20	31.52%	22.23%	20.31%	18.71%	18.69%	18.66%
	40	31.49%	22.15%	20.25%	18.58%	18.48%	18.45%
	60	31.48%	22.12%	20.23%	18.53%	18.34%	18.31%
	100	31.38%	22.01%	20.12%	18.44%	17.67%	17.64%
fr-en	1	31.34%	23.74%	20.08%	20.89%	21.15%	21.00%
	5	31.27%	21.93%	20.04%	19.12%	19.14%	19.09%
	10	31.20%	21.84%	19.95%	18.59%	18.60%	18.56%
	20	31.21%	21.80%	19.86%	18.48%	18.37%	18.34%
	40	31.22%	21.78%	19.87%	18.41%	18.23%	18.19%
	60	31.20%	21.72%	19.84%	18.34%	18.13%	18.09%
	100	31.10%	21.65%	19.76%	18.24%	17.83%	17.80%
es-pt	1	31.15%	23.52%	20.43%	21.08%	20.19%	19.55%
	5	31.23%	22.28%	20.35%	19.43%	18.14%	17.79%
	10	31.15%	22.21%	20.18%	19.14%	17.65%	17.34%
	20	31.16%	22.09%	20.14%	18.95%	17.34%	17.08%
	40	31.16%	22.05%	20.12%	18.82%	17.25%	17.00%
	60	31.16%	22.01%	20.10%	18.76%	17.18%	16.93%
	100	31.10%	21.94%	20.02%	18.70%	17.08%	16.83%