Alibi corpus documentation

Joanna Radoła

June 2024

Contents

1 General information

The corpus consists of five bitexts consisting of a French short story and its translation into English. A translator-made hierarchical alignment of these files

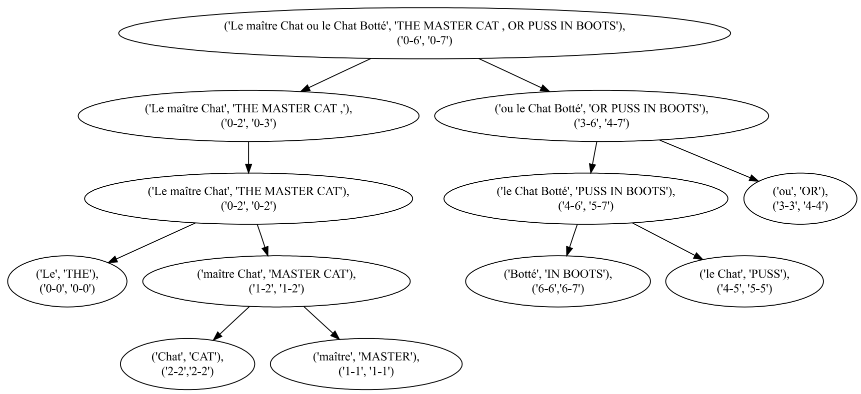

Figure 1: A visualization of the hierarchical alignments

is also provided, relying on a methodology described in [2].[1]. These bitexts are the following (for each, we also use a one-letter code, e.g., (A) for *l'Auberge*):

1. (A) *L'Auberge* - *The Inn*, from *Le Horla* by Guy de Maupassant,

2. (B) *La Barbe Bleue* - *The Blue Beard*, from *Les Contes de ma mère l'Oye* by Charles Perrault,

3. (C) *Le Chat Botté* - *Master Cat*, from *Les Contes de ma mère l'Oye* by Charles Perrault,

4. (D) *La Dernière Classe* - *The Last Lesson*, from *Les contes du lundi* by Alphonse Daudet,

5. (V) *Vision de Charles XI* - *The Vision of Charles XI, from Colomba et autres contes et nouvelles* by Prosper Mérimée.

A sample hierarchical alignment tree is shown in 1. Aligned spans in French and English are represented as strings and as ranges of word indexes (ids). The first word in a sentence is at index 0.

[1] See also the associated guidelines

The hierarchical alignments from the XML files have been converted into the so-called "extended Pharaoh format":

- The tokenized source and target texts are stored in separate files, one line per sentence. Each line starts with a unique sentence id, followed by a `TAB`, followed by a space-separated list of tokens.
- Each line of the alignment file also corresponds to one sentence. It begins with a sentence id, followed by a `TAB`, and a space-separated list of links.
- The links are of form $n-m$ for *sure links* and npm for *possible links* (where n and m are the word indexes, starting at index 0).

An example is on Figure 2.

```
ChatBotte.txt
---------------------------------------------------------------------
0   Le maître Chat ou le Chat Botté
1   Ils auraient eu bientôt mangé tout le pauvre patrimoine .
2   voilà monsieur le marquis de Carabas qui se noie ! "
---------------------------------------------------------------------
MasterCat.txt
---------------------------------------------------------------------
0   THE MASTER CAT , OR PUSS IN BOOTS
1   They would soon have eaten up all the poor property .
2   My Lord the Marquis of Carabas is drowning ! "
---------------------------------------------------------------------
ChatBotte_MasterCat.ali.txt
---------------------------------------------------------------------
0   0-0 1-1 2-2 3-4 4p5 5p5 6p6 6p7
1   0-0 1-1 2-3 3-2 4p4 4p5 5-6 6-7 7-8 8-9 9-10
2   1p0 1p1 2-2 3-3 4-4 5-5 7p6 7p7 8p6 8p7 9-8 10-9
```

Figure 2: An example of the Pharaoh format

The heuristics used to convert spans into word-to-word alignments use the following rules:

1. a leaf node corresponding to a one-to-one alignment is turned into a sure link;
2. for every word in one language in an aligned leaf (span that hasn't been divided any further), draw a possible link to all the words of the leaf in the other language. For example, a leaf alignment *La Barbe-Bleue, Blue Beard* is converted to *0p0 0p1 1p0 1p1*.

2 Corpus Statistics

2.1 Number of sentences, of tokens

The French and the English sentences have been prealigned and tokenized with Europarl tools. Word-level tokenization is used throughout the project, with punctuation marks treated as separate tokens.

text_id	sentences	fr tokens	en tokens
A	203	5674	5862
B	67	2145	2234
C	51	1861	2036
D	91	2045	1935
V	105	2912	2790
total	517	14637	14857

Table 1: Number of sentences, of French and English tokens in each bitext.

2.2 Number of sure/possible alignments, of all spans, of leaf spans

text_id	all spans	leaf spans	sure ali	possible ali
A	9144	4368	3389	6102
B	3311	1522	1151	1913
C	3091	1453	1173	1416
D	2902	1420	1058	2174
V	4480	2165	1681	2542
total	22928	10928	8452	14147

Table 2: Number of spans obtained from recursive divisions of every sentence, number of leaf spans (those that were not divided further), the number of sure and possible alignments obtained from leaves.

2.3 Distribution of lengths of leaves

2.3.1 French-English bitext leaves

lengths	number of leaves
1-1	8452
2-1	643
1-2	632
2-2	407
2-3	108
1-3	101
3-2	97
3-3	86
3-1	76
3-4	34
4-1	32
4-3	31
1-4	29
4-2	26
2-4	21
4-4	19
other	134
total	10928

Table 3: The number of bitext leaves with lengths $n-m$ (French span of length n, English span of length m) in all bitexts combined.

2.3.2 French and English texts separately

	A		B		C		D		V		total	
nb	fr	en	fr	en	fr	en	fr	en	fr	en	fr	en
1	3724	3694	1246	1266	1271	1244	1146	1168	1828	1836	9215	9208
2	476	461	194	175	119	145	155	148	242	242	1186	1171
3	112	140	44	56	36	33	60	61	54	54	306	344
4	36	55	18	8	18	18	33	23	24	18	129	122
5	9	7	7	8	4	5	16	11	8	6	44	37
6	4	3	7	7	3	4	6	4	3	1	23	19
7	1	3	4	1	2	2	3	2	2	1	12	9
8	1	1	2	0	0	2	1	3	0	4	4	10
9	1	1	0	1	0	0	0	0	1	1	2	3
10+	4	3	0	0	0	0	0	0	2	2	6	5

Table 4: The number of leaves which have *nb* words for French and for English in texts A, B, C, D, V.

2.4 Aligned/unaligned words

	A		B		C		D		V		total	
-	fr	en	fr	en	fr	en	fr	en	fr	en	fr	en
ali wo	5338	5430	1959	1914	1741	1784	1913	1856	2712	2672	13663	13656
wo total	5674	5862	2145	2234	1861	2036	2045	1935	2912	2790	14637	14857
non-ali	336	432	186	320	120	252	132	79	200	118	974	1201

Table 5: The number of words included in leaves, the total number of words, the difference between the two.

2.5 POS-tagging results

The tokens were tagged with the use of the Stanford POS Tagger. The tagsets are presented in figures 3 and 4.

ADJ	adjectif
ADJWH	adjectif interrogatif
ADV	adverbe
ADVWH	adverbe interrogatif
CC	conjonction de coordination
CL	pronom clitique
CLO	pronom clitique objet
CLR	pronom clitique réfléchi
CLS	pronom clitique sujet
CS	conjonction de subordination
DET	déterminant
DETWH	déterminant interrogatif
ET	mot tiré d'une langue étrangère
I	interjection
NC	nom commun
NPP	nom propre
P	préposition
P+D	forme contractée préposition et déterminant
P+PRO	forme contractée préposition et pronom
PONCT	ponctuation
PREF	préfixe
PRO	pronom
PROREL	pronom relatif
PROWH	pronom interrogatif
V	verbe
VIMP	forme verbale à l'impératif
VINF	forme verbale à l'infinitif
VPP	participe passé
VPR	participe présent
VS	forme verbale au subjonctif

Figure 3: French POS tagset.

Table 2
The Penn Treebank POS tagset.

1.	CC	Coordinating conjunction	25.	TO	*to*
2.	CD	Cardinal number	26.	UH	Interjection
3.	DT	Determiner	27.	VB	Verb, base form
4.	EX	Existential *there*	28.	VBD	Verb, past tense
5.	FW	Foreign word	29.	VBG	Verb, gerund/present participle
6.	IN	Preposition/subordinating conjunction	30.	VBN	Verb, past participle
7.	JJ	Adjective	31.	VBP	Verb, non-3rd ps. sing. present
8.	JJR	Adjective, comparative	32.	VBZ	Verb, 3rd ps. sing. present
9.	JJS	Adjective, superlative	33.	WDT	*wh*-determiner
10.	LS	List item marker	34.	WP	*wh*-pronoun
11.	MD	Modal	35.	WP$	Possessive *wh*-pronoun
12.	NN	Noun, singular or mass	36.	WRB	*wh*-adverb
13.	NNS	Noun, plural	37.	#	Pound sign
14.	NNP	Proper noun, singular	38.	$	Dollar sign
15.	NNPS	Proper noun, plural	39.	.	Sentence-final punctuation
16.	PDT	Predeterminer	40.	,	Comma
17.	POS	Possessive ending	41.	:	Colon, semi-colon
18.	PRP	Personal pronoun	42.	(	Left bracket character
19.	PP$	Possessive pronoun	43.	)	Right bracket character
20.	RB	Adverb	44.	"	Straight double quote
21.	RBR	Adverb, comparative	45.	‘	Left open single quote
22.	RBS	Adverb, superlative	46.	“	Left open double quote
23.	RP	Particle	47.	’	Right close single quote
24.	SYM	Symbol (mathematical or scientific)	48.	”	Right close double quote

Figure 4: English POS tagset.

2.5.1 Most frequently aligned POS pairs

The highest frequencies are observed for pairs of common nouns and functional words, specifically prepositions and determiners. Perhaps unsurprisingly, another frequent pair contains punctuation.

#	fr-en tag pair	count
1	NC-NN	1462
2	P-IN	1060
3	DET-DT	1045
4	PONCT-”	871
5	V-VBD	850
6	CLS-PRP	508
7	ADV-RB	492
8	ADJ-JJ	489
9	NC-NNS	486
10	PONCT-.	480
11	CC-CC	362
12	DET-PRP$	338
13	P-DT	313
14	DET-NN	296
15	P-NN	262
16	P-TO	255
17	NC-IN	240
18	VINF-VB	236
19	NC-DT	226
20	NPP-NNP	222
21	DET-IN	200
22	NC-NNP	198
23	VPP-VBN	180
24	CLO-PRP	173
25	NC-JJ	172
26	P-RB	168
27	V-RB	162
28	CS-IN	154
29	V-VB	151
30	ADV-IN	150

Table 6: 30 most frequent POS tag pairs (out of total 606 pairs).

2.5.2 Frequency of POS tags in French and English texts separately

The most frequent tags in both languages are nouns, prepositions, determiners, and punctuation.

POS	A	B	C	D	V	all
NC	991	307	294	322	541	2455
DET	770	263	235	239	425	1932
PONCT	716	273	214	243	335	1781
P	702	214	194	236	370	1716
V	487	221	199	190	248	1345
ADJ	383	95	89	115	177	859
ADV	278	131	78	149	137	773
CLS	213	125	83	117	79	617
CC	210	75	60	52	83	480
VINF	147	68	73	78	69	435
NPP	158	46	86	47	94	431
VPP	145	46	44	50	77	362
CLO	61	73	53	62	37	286
CS	82	59	43	44	48	276
PROREL	70	36	35	25	42	208
CLR	101	28	16	16	33	194
PRO	59	45	21	32	36	193
ET	32	12	18	13	45	120
VPR	57	17	18	6	16	114
VS	3	2	3	1	7	16
CL	0	1	0	1	5	7
PROWH	0	1	1	2	3	7
ADVWH	3	2	1	0	0	6
I	1	0	0	3	0	4
PREF	0	2	1	0	0	3
ADJWH	0	0	0	1	1	2
DETWH	0	0	0	0	2	2
VIMP	0	0	1	0	0	1
sum	5669	2142	1860	2044	2910	14625

Table 7: French POS tags in all texts.

POS	A	B	C	D	V	all
NN	710	252	201	201	376	1740
IN	617	182	191	182	327	1499
DT	565	168	204	154	340	1431
,	480	191	167	134	174	1146
VBD	481	150	148	137	174	1090
PRP	325	181	154	163	108	931
JJ	362	89	71	99	167	788
RB	294	113	86	129	110	732
CC	276	85	69	66	83	579
.	203	80	64	93	127	567
NNS	228	71	60	66	112	537
NNP	163	70	85	45	117	480
VB	152	88	79	76	72	467
PRP$	145	73	66	65	75	424
TO	156	61	69	57	56	399
VBN	143	33	45	39	98	358
VBG	127	35	33	41	32	268
”	23	48	37	17	32	157
RP	81	26	14	20	15	156
MD	28	27	22	24	21	122
VBP	22	31	28	18	21	120
“	22	42	20	8	23	115
WDT	56	17	13	7	21	114
CD	60	22	8	6	15	111
:	23	28	16	23	16	106
VBZ	27	18	12	11	25	93
WP	20	10	27	6	18	81
WRB	24	10	8	22	5	69
PDT	11	13	7	4	4	39
POS	9	3	7	3	4	26
EX	7	3	2	4	5	21
RBR	10	2	2	3	4	21
JJS	1	6	6	0	2	15
NNPS	2	0	9	2	1	14
UH	4	3	2	3	2	14
JJR	2	1	2	4	4	13
RBS	0	2	1	3	3	9
WP$	3	0	0	0	1	4
LS	0	0	1	0	0	1
sum	5862	2234	2036	1935	2790	14857

Table 8: English POS tags in all texts.

3 Baseline alignment scores

For these baselines we use SimAlign [1], a tool leveraging multilingual word embeddings to automatically generate word alignments without the necessity to pre-train it on parallel data. We use simalign with the following parameters (`model`='xlmr', `layer`=8, `token_type`='word'). After building a matrix of cosine similarities between tokens in the source and the target, it uses basic three methods to compute alignment links:

- Argmax - an alignment $n-m$ is made iff word n has the highest similarity to m and vice versa.

text	precision	recall	F1	AER
A	0.957	0.881	0.917	0.077
B	0.937	0.899	0.918	0.079
C	0.958	0.922	0.94	0.058
D	0.943	0.869	0.904	0.089
V	0.947	0.88	0.912	0.083

Table 9: Results obtained with the Argmax method. The alignments provided by a human translator were hereafter taken as the reference.

- Itermax - 2 or more iterations of the argmax method.

text	precision	recall	F1	AER
A	0.907	0.937	0.922	0.081
B	0.887	0.943	0.914	0.091
C	0.904	0.961	0.932	0.072
D	0.898	0.915	0.906	0.096
V	0.904	0.933	0.918	0.084

Table 10: Results obtained with the Itermax method.

- Match - maximum-weight matching in the bipartite graph induced by the similarity matrix.

text	precision	recall	F1	AER
A	0.843	0.961	0.898	0.112
B	0.807	0.955	0.875	0.14
C	0.846	0.971	0.904	0.105
D	0.846	0.95	0.895	0.117
V	0.832	0.942	0.884	0.126

Table 11: Results obtained with the Match method.

It can be observed that while Argmax is the method that yields the lowest AER, Itermax outperforms it in F-score in three out of five cases. The two methods illustrate a tradeoff between high precision and high recall - Argmax is very selective and therefore tends to make correct predictions, but also to make fewer of them, which results in lower recall. Applying Itermax improves the recall, as more correct alignments are predicted, however, it comes at a cost of more incorrect predictions and precision drops. Match method scores the highest on recall. As usual, which method is the most suitable depends on the use case.

4 Typos, omissions

4.1 Ids of leaves longer than 10 tokens (weren't divided further)

4.1.1 *L'Auberge/The Inn*

1. 21.1_0-27_0-25
2. 96.1_0-26_0-31
3. 192.1_4-27_9-33

4.1.2 *La Vision/The Vision*

1. 17.2_13-31_11-18
2. 70.2_5-22_5-16
3. 99.2_7-17_7-17

4.2 Sentences in *La Vision/The Vision* that have tokenization/spelling errors

1. linkGroup id=”60” stop ping
2. linkGroup id=”61” gal- lery
3. linkGroup id=”72” fiUed
4. linkGroup id=”93” Chahles
5. linkGroup id=”96” after ward
6. linkGroup id=”98” aU
7. linkGroup id=”100” pres- ent
8. linkGroup id=”103” wiU circimi- stances

5 Appendix

5.1 Additional corpus statistics

text_id	fr/en words	fr words/sentence	en words/sentence
A	0.968	27.951	28.877
B	0.960	32.015	33.343
C	0.914	36.490	39.922
D	1.057	22.473	21.264
V	1.044	27.733	26.571
all texts	0.985	28.311	28.737

Table 12: The ratio of French words to English words and the mean number of words per sentence for both languages.

text_id	fr words/span	en words/span	fr words/leaf	en words/leaf
A	3.994	4.160	1.222	1.243
B	4.205	4.256	1.287	1.258
C	3.952	4.217	1.198	1.228
D	4.024	3.858	1.347	1.307
V	4.136	4.010	1.247	1.229
all texts	4.062	4.100	1.260	1.253

Table 13: The mean number of words per span (including leaf spans) and mean number of words per leaf span in both languages.

References

[1] Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[2] Yong Xu and François Yvon. Novel elicitation and annotation schemes for sentential and sub-sentential alignments of bitexts. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth Language Resources and Evaluation Conference (LREC 2016)*, page 10, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).