Adding Syntactic Annotations to Flickr30k Entities Corpus for Multimodal Ambiguous Prepositional-Phrase Attachment Resolution

Sebastien Delecraz, Alexis Nasr, Frederic Bechet, Benoit Favre
Aix-Marseille Univ, Université de Toulon, CNRS, LIS
163 avenue de Luminy - Case 901
F-13288 Marseille - Cedex 9 - FRANCE
{firstname.lastname}@univ-amu.fr

Abstract

We propose in this paper to add to the captions of the *Flickr30k Entities* corpus some syntactic annotations in order to study the joint processing of image and language features for the Preposition-Phrase attachment disambiguation task. The annotation has been performed on the English version of the captions and automatically projected on their French and German translations.

Keywords: PP-attachments, Multimodal Corpus, Multilingual Corpus

1. Introduction

Joint processing of image and text is a very active area of research. It is studied mostly in the context of natural language generation, for example for generating a textual description of a video or an image. Recently open-domain language generation from images or videos received a lot of attention through the use of multimodal deep neural networks (Vinyals et al., 2015). Theses models build a unified representation for both image and language features and generate in an end-to-end process a text directly from an image, without an explicit representation (syntactic or semantic) of the text generated.

In this paper we propose to study joint image and language processing for language parsing rather than generation. The main idea of this study is to check if the use of visual features extracted from an image can be useful in order to disambiguate the linguistic analysis of a caption that describes the same image. To do so we propose a new framework for testing multimodal approaches on the specific task of ambiguous Prepositional-Phrase attachment (*PP-attachment*) resolution.

PP-attachments are known to be an important source of errors in parsing natural language. The main reason being that, in many cases, correct attachments cannot be predicted accurately based on pure syntactic considerations: their prediction ask for precise lexical co-occurrences. Such information is usually not found in treebanks that are limited in their size and therefore do not model many bi-lexical phenomena. Besides, disambiguation may ask for non linguistic knowledge which is not present in treebanks.

In this paper, we propose to create a corpus for supporting PP-attachment disambiguation research by combining textual and visual information. The contribution of this study is the selection and the manual annotation of a corpus of ambiguous PP-attachments from the multimodal corpus *Flickr30k Entities* (Plummer et al., 2017). A full parse of the sentences containing a hand-corrected PP-attachment, which is compatible with the manual attachment is also produced. In addition, we use MT alignments to transfer the annotations to French and German translations from the *Multi30k* corpus (Elliott et al., 2016; Elliott et al., 2017).

1. someone is holding out a punctured ball in front of a brown dog with a red collar .
2. A man holding out a deflated soccer ball to a gray dog .
3. The owner tries to hand a deflated ball to his dog .
4. Large gray dog being handed a white soccer ball .
5. A brown dog starring at a soccer ball .

Figure 1: Example of the *F30kE* annotations. The image is described with five captions, each annotated with entities. Entities that corefer with a visual element in the image are linked to the corresponding bounding box.

Finally, for every preposition manually attached, a set of possible attachment alternatives for use in a reranking system is produced.

2. Enriching the Flickr30k Entities Corpus with PP-Attachment Annotations

Corpora with joint annotation of image and text has recently become widely available. The corpus used in this work is the *Flickr30k Entities* (*F30kE*) (Plummer et al., 2017), an extension of the original *Flickr30k* dataset (Young et al., 2014). This corpus is composed of almost $32K$ images and, for each image, five captions describing the image have been produced. Besides, every object in the image that corresponds to a mention in the captions has been manually identified with a bounding box.

Preposition	Occ.	% Noun	% Verb	Dist.
in	4191	0.59	0.41	2.21
with	3018	0.59	0.41	2.47
for	1777	0.36	0.64	1.57
near	1452	0.65	0.35	1.69
through	1420	0.05	0.95	2.01
on	1359	0.26	0.74	2.03
next to	1342	0.08	0.92	3.23
from	1172	0.30	0.70	2.43
into	1123	0.08	0.92	2.32
over	941	0.39	0.61	2.38
by	890	0.10	0.90	2.60
at	720	0.20	0.80	2.75
of	700	0.97	0.03	1.03
around	589	0.18	0.82	2.26
in front of	570	0.11	0.89	3.72
under	544	0.18	0.82	3.29
behind	544	0.35	0.65	1.78
along	500	0.37	0.63	1.79
during	423	0.14	0.86	5.08
across	415	0.11	0.89	2.21
down	393	0.66	0.34	2.53
against	365	0.39	0.61	1.56
outside	356	0.38	0.62	2.28
towards	276	0.08	0.92	2.41
out of	252	0.13	0.87	2.08
beside	245	0.03	0.97	3.51
above	241	0.43	0.57	2.90
in the middle of	240	0.12	0.88	3.63
onto	210	0.08	0.92	2.34
outside of	206	0.12	0.88	3.30
inside	197	0.13	0.87	3.72
between	189	0.72	0.28	1.29
past	170	0.19	0.81	2.95
toward	167	0.34	0.66	1.53
on top of	166	0.14	0.86	2.93
like	159	0.34	0.66	1.32
among	142	0.37	0.63	2.13
after	126	0.10	0.90	3.12
away from	109	0.04	0.96	1.75
off	104	0.31	0.69	3.72
up	96	0.82	0.18	1.93
up to	86	0.12	0.88	2.35
before	71	0.10	0.90	4.14
atop	60	0.23	0.77	2.97
about	54	0.59	0.41	1.52
along with	54	0.15	0.85	4.98
underneath	49	0.18	0.82	3.16
without	46	0.54	0.46	2.98
out	42	0.29	0.71	4.43
at the top of	40	0.10	0.90	2.23
inside of	39	0.23	0.77	3.72
amongst	33	0.09	0.91	3.18
close to	33	0.15	0.85	4.42
upon	31	0.13	0.87	1.71
amidst	28	0.29	0.71	3.18
beneath	26	0.12	0.88	3.54
within	24	0.33	0.67	3.62
below	23	0.52	0.48	1.83
at the bottom of	22	0.18	0.82	2.86
amid	22	0.18	0.82	5.00
in between	18	0.22	0.78	2.89
up against	16	0.06	0.94	3.69
ahead of	14	0.00	1.00	1.71
together with	13	0.08	0.92	2.54
such as	13	0.38	0.62	5.31
besides	12	0.42	0.58	2.33
beyond	10	0.70	0.30	3.30
on the top of	10	0.10	0.90	2.70
while	10	0.20	0.80	6.40
near to	10	0.10	0.90	4.10
Total	29068	0.36	0.64	2.37

Table 1: Prepositions annotated with their occurrence number in the corpus and statistics about their attachment.

Bounding boxes and the mentions in the captions have been paired together via coreference links. A total of $244K$ such links have been annotated.

Furthermore, each mention in the captions has been categorized into eight coarse-grained conceptual types. These types are: people, body parts, animals, clothing, instruments, vehicles, scene, and other. One example of the corpus has been reproduced in Figure 1.

Captions in the *F30kE* corpus are annotated at the conceptual level, but no syntactic annotation is provided. Since our goal in this study is to evaluate several sets of multimodal features for the PP-attachment task, we needed to add such a level of annotation to the corpus.

We did not have the resources for manually annotating the whole *F30kE* caption corpus with syntactic annotations. Therefore we limited our effort to the manual annotation of PP-attachments in ambiguous contexts.

In order to select ambiguous PP-attachments we applied the following process: first the captions of *F30kE* were processed by a Part-Of-Speech tagger (Nasr et al., 2011); then a set of regular expressions on the POS labels were defined in order to select sentences that contain a preposition that can be attachment to more than one word; finally, the ambiguous prepositions have been manually attached to their correct syntactic governor.

Captions containing ambiguous PP-attachment have been identified using two simple rules: a preposition is considered ambiguous if it is preceded by at least two nouns or a verb and a noun, in other word, the captions must match one of the following regular expressions:

- `X* N X* N X* p X*`
- `X* V X* N X* p X*`

where `N` and `V` stand for the POS tags noun and verb, `X` stand for any POS tag and `p` is the target preposition. $22,800$ captions were selected this way, that correspond to $15,700$ different images. They constitute our *PP-Flickr* corpus. This corpus contains $29,068$ preposition occurrences that have been manually attached to their syntactic governor. In the manual annotation process information given to the annotator is limited to a caption, the target preposition which needs to be attach to its governor, and the corresponding image.

Table 1 presents an overview of some statistics on the corpus. For each preposition, ordered by occurrence, we give the rate of attachment to a noun and a verb and the average distance between the preposition and its governor. For readability, we did not display the 25 prepositions with less than 10 occurrences but the total counts are computed on the whole corpus. 64% of the prepositions occurrences are attach to a verb and the average distance between the preposition and its governor is 2.37 words. The ten more frequent prepositions represent 61.22% of the annotated corpus. For preposition with a highest average distance ($>$ 3 words) that represent 25 prepositions (*i.e.* 4497 attachments), the governor is a verb in 74% of cases.

3. A Multimodal Corpus for Syntactic Parse Reranking

The annotation process described in the previous section only provides links between a preposition and its governor. We added syntactic annotations to the corpus in order to propose a task of multimodal reranking of syntactic parsing hypotheses on the *F30kE* corpus.

The first step in this process was to parse all the captions manually annotated with PP-attachment with a transition-based dependency parser. The parser allows to set some dependencies, prior to parsing, and generate the most likely parse including these dependencies. In our case all the PP-attachments manually annotated were set, then the parser provided the best dependency parse compatible with these attachments. Although these automatic annotations cannot be considered as *gold* labels, they can be considered as fairly robust as the main source of ambiguities (and therefore errors) is neutralized.

Once the best syntactic parse for each caption of the corpus is produced, we apply a method for generating alternative PP-attachment sites: given a sentence S, a parse T for S and a target preposition p, we define a set G_p of candidate governors for p. The set G_p is initialized with g, the actual governor of p in the parse T. The following rules are then applied to T and new potential governors are added to G_p:

1. $N \leftarrow V \rightarrow p \quad \Rightarrow G_p = G_p \cup \{N\}$
2. $N \leftarrow P \leftarrow V \rightarrow p \quad \Rightarrow G_p = G_p \cup \{N\}$
3. $N' \leftarrow N \rightarrow p \quad \Rightarrow G_p = G_p \cup \{N'\}$
4. $N' \leftarrow P \leftarrow N \rightarrow p \quad \Rightarrow G_p = G_p \cup \{N'\}$
5. $N' \rightarrow X \rightarrow N \rightarrow p \quad \Rightarrow G_p = G_p \cup \{N'\}$
6. $N \rightarrow N \rightarrow p \quad \Rightarrow G_p = G_p \cup \{N\}$
7. $V \rightarrow N \rightarrow p \quad \Rightarrow G_p = G_p \cup \{V\}$

These rules are inspired by the ideas of (Anguiano and Candito, 2011; Attardi and Ciaramita, 2007; Hall and Novák, 2005). For example, the rule 1 is interpreted as follows: if target preposition p has a verbal governor which has a noun N as a direct dependent, N is added as a candidate governor. In rule 2, if the target preposition p is dependent of a verb V, it can also be attached to any noun that is itself governed by another preposition attached to V. The application of all rules must meet a general condition which is that the tree produced must be projective. These rules have been designed in such a way that most possible governors are included in the set G_p. The application of these rules on the test set showed that in 92.28% on the cases, the correct governor is in G_p.

Given the sentence *a man throws a child into the air at a beach*, and target preposition *at* that the parser has attached to *child*, the two rules 4 and 7 apply, yielding $G_p = \{child,\ air,\ throws\}$

	man	*throws*	*child*	*into*	*air*	*at*
4.			N	P	$N*$	p
7.		$V*$	N			p

Thanks to this process we have now for each caption with an ambiguous PP-attachment a set of syntactic parses that differ only by the governor chosen for the target preposition. We can now introduce the task of *multimodal syntactic parsing reranking* that can take advantage of all the visual features available in the *F30kE* corpus for finding the correct PP-attachment among all the possible parses.

4. Baseline PP-attachment reranking

In order to calibrate future research, we provide baseline results for PP-attachment reranking. The task consists in selecting the correct attachment from a list of potential attachments. A classifier is first trained to detect incorrect PP-attachments, and then the highest scoring PP-attachment alternative is output.

The classifier used as baseline is the *Icsiboost* classifier (Favre et al., 2007). This *Adaboost* classifier is a combination of weak learners that learn a threshold for continuous features, and a binary indicator for discrete ones. Training minimizes the exponential loss function by greedily selecting the best classifier and re-weighing the training set to focus on misclassified examples. This classifier is a strong baseline as it performs feature selection and has been shown to perform well on a range of tasks.

The features used to train the classifiers are defined for a governor-preposition-dependent triplet.

- (P)reposition: lemma of preposition
- (T)ext: part-of-speech and lemma of governor, dependent, both, and syntactic function of preposition, distance between governor and preposition.
- Visual (C)oncepts: concept of governor, concept of dependent, concepts of both
- Visual (S)patial: the normalized distance between the top-left and bottom-right corners of the governor and dependent bounding boxes, the areas of those boxes, and the ratio between the areas of the governor and dependent boxes.

It is important to notice that the visual features in our study are limited to spatial information about bounding boxes and visual concepts. No image analysis of the content of the boxes is done since this level of information is covered by the *visual concept features* which attach to each box a concept tag related to its content.

Table 2 presents the accuracy of PP-attachment after correction according to the candidates provided in the corpus, with different feature set combinations. Adding conceptual features to textual features improves accuracy, however spatial features have no impact when used in conjunction with other feature sets.

Features	Test
Baseline	0.75
P+T	0.85
P+C	0.82
P+S	0.77
P+T+C	0.86
P+T+S	0.86
P+C+S	0.82
P+T+C+S	0.86

Table 2: Baseline PP-attachment reranking accuracy on the test set.

5. Multilingual Extension

We extend the corpus by taking advantage of the translations produced for the WMT shared task on multimodal machine translation (Elliott et al., 2017). That corpus contains $31,014$ German (Elliott et al., 2016) and French translations from the *Flickr30k* corpus created by professional translators (German) and crowd sourcing (French). One arbitrary caption was translated per image to both languages. Out of those $31K$ translations, $5,225$ captions overlap with the gold standard PP attachment annotations.

In order to transfer the annotations across languages, we align them with the `fast_align` program (Dyer et al., 2013) and merge forward and reverse alignments with the `grow-diag-final-and` heuristic. In order to improve the accuracy of the alignment, we concatenate the translated sentences from the *F30kE* corpus with the news commentary *Bitexts* and the *Freedict* bilingual dictionary. This process helps aligning common expressions by providing more evidence to the unsupervised algorithm. From the alignments, we propagate two types of informations. First, the *Flickr30k Entities* segments which include unique entity references and types. This propagation is available for the whole set of translated captions, and could be used for instance to train a phrase retrieval system in the images, a coreference tracking system or other type of systems exploiting the *Flickr30k Entities* data, but in French and German. The second type of information propagated is the PP-attachment gold standard. The mapping is performed by transferring the annotation of preposition, and the head word of the governing entity and the hypothesis generated by the baseline English parser, to the words they are aligned to in the target language. If the source language words are aligned to multiple words, we use the first word by word order in the sentence.

The resulting multilingual annotations are made available along with the rest of the corpus in order to foster parsing research in all of the three languages, as well as the interesting link between PP-attachment resolution and machine translation. The transfer results in $30K$ sentences with entity boundaries, types and identity in French and German, as well as $5,225$ sentences with gold PP attachment hypotheses and gold standard in those languages (Figure 2).

The quality of transfer is highly dependent on the quality of the automatic alignment, a known difficult problem, for which the error rate is typically around 30%. We analyzed a random sample of 100 alignments for the French subset,

EN Large furry dog [G walking] in the [H sand] [P near] large rocks .

FR Un gros chien poilu [G marchant] dans le [H sable] [P près de] gros rochers .

DE Großer Hund mit langem Fell [G läuft] in der [P Nähe] großer Felsen durch den [H Sand] .

Figure 2: Example of annotations transferred from English to French and German. [P] represents the target preposition, [H] is the baseline governor predicted by the parser for English, [G] represents the gold standard manually annotated. Note that entities and coreference links are also transferred while not depicted in this figure.

and manually annotated incorrect propagations. The errors can come from the preposition, its gold governor, or predicted governor not being aligned, or being aligned to a word with the wrong part-of-speech. Often multiword idiomatic expressions in either language, that are characterized by arbitrary head-words, usually result in wrong alignment (*to bike* $\rightarrow$ *faire du vélo*, where *bike* is aligned with *vélo*, which might also be due to the fact that *bike* can both be a verb and a noun). Another frequent problem is the preposition being aligned to the wrong preposition in the target language when the target sentence contains several prepositions.

The result of this hand analysis is that 21% of transfers in the sample are erroneous. The transfer quality could be improved by accounting for part of speech tags, for example by using a joint grammar of both languages to enforce constituent-level alignments. A better heuristic could also be devised for processing multiply aligned source words for which the choice of alignment is not always successful. Finally, enforcing that aligned entities should have the same semantic category could improve the confidence of the transfer. We have released all the tools used to generate the transferred annotations in hope that they can be extended to improve the final result.

6. Distributed Data

The annotated corpus is available at `https://gitlab.lis-lab.fr/sebastien.delecraz/pp-flickr.git`. The annotation are given in JSON format for the three languages (English, French and German). In addition, we provide the English corpus in

#	Word	Lemma		POS		Governor	Label	Entity ID	Gold Gov.
1	someone	someone	_	NN	_	2	SBJ	227018	_
2	is	be	_	VBZ	_	0	ROOT	_	_
3	holding	hold	_	VBG	_	2	VC	_	_
4	out	out	_	RP	_	3	PRT	_	_
5	a	a	_	DT	_	7	NMOD	227019	_
6	punctured	puncture	_	VBN	_	7	NMOD	227019	_
7	ball	ball	_	NN	_	3	OBJ	227019	_
8	in front of	in front of	_	IN	_	3	ADV	_	3
9	a	a	_	DT	_	11	NMOD	227017	_
10	brown	brown	_	JJ	_	11	NMOD	227017	_
11	dog	dog	_	NN	_	8	PMOD	227017	_
12	with	with	_	IN	_	11	NMOD	_	11
13	a	a	_	DT	_	15	NMOD	227021	_
14	red	red	_	JJ	_	15	NMOD	227021	_
15	collar	collar	_	NN	_	12	PMOD	227021	_
16	.	.	_	.	_	2	P	_	_

Table 3: Example an annotated sentence in CoNLL format

CoNLL format (Table 3). Columns one to eight correspond to the standard CONLL columns, column nine corresponds to entity ID in the *F30kE* and column ten indicates whether the dependency is hand-corrected, such as words 8 and 12, or not, which is the case for all other words.

7. Conclusion

We have proposed in this paper a corpus for supporting PP-attachment reranking research when attachments can be disambiguated with an image. The corpus was created by enriching the *Flickr30k Entities* corpus with $29,068$ PP-attachments, from $22,800$ captions describing $15,700$ images, manually resolved by looking at the images.
We provide a testbed for reranking attachments generated from a forced parse with the correct attachment and a set of rules. A baseline classifier using reference visual features (concepts and spatial relations), and textual features yields PP-attachment accuracy of 86% from an original accuracy of 75% form a standard transition-based parser. The corpus is enriched with multilingual annotations transfered to French and German through automatic alignment.

8. Acknowledgements

This work has been carried out thanks to the support of French DGA in partnership with Aix-Marseille University as part of the "Club des partenaires Défense".

9. Bibliographical References

Anguiano, E. H. and Candito, M. (2011). Parse correction with specialized models for difficult attachment types. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1222–1233. Association for Computational Linguistics.

Attardi, G. and Ciaramita, M. (2007). Tree revision learning for dependency parsing. In *HLT-NAACL*, pages 388–395.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. Association for Computational Linguistics.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017). Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark, September.

Favre, B., Hakkani-Tür, D., and Cuendet, S. (2007). Icsiboost. `http://code.google.come/p/icsiboost`.

Hall, K. and Novák, V. (2005). Corrective modeling for non-projective dependency parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 42–52. Association for Computational Linguistics.

Nasr, A., Béchet, F., Rey, J., Favre, B., and Le Roux, J. (2011). Macaon: An nlp tool suite for processing word lattices. *Proceedings of the ACL 2011 System Demonstration*, pages 86–91.

Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2017). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123(1):74–93.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.