Disambiguation of English PP Attachment using Multilingual Aligned Data

Lee Schwartz, Takako Aikawa, Chris Quirk
Microsoft Research
One Microsoft Way
Redmond, WA 98008,
{leesc, takakoa, chrisq} @microsoft.com

Abstract

Prepositional phrase attachment (PP attachment) is a major source of ambiguity in English. It poses a substantial challenge to Machine Translation (MT) between English and languages that are not characterized by PP attachment ambiguity. In this paper we present an unsupervised, bilingual, corpus-based approach to the resolution of English PP attachment ambiguity. As data we use aligned linguistic representations of the English and Japanese sentences from a large parallel corpus of technical texts. The premise of our approach is that with large aligned, parsed, bilingual (or multilingual) corpora, languages can learn non-trivial linguistic information from one another with high accuracy. We contend that our approach can be extended to linguistic phenomena other than PP attachment.

1 Introduction

English is syntactically ambiguous with respect to PP attachment. For instance, the PP in (1) can be attached either to an NP as in (1a) or to a VP as in (1b).

(1) Drag the file next to the item.
a. *NP attachment*:
Drag [$_{NP}$ the file [$_{PP}$ next to the item]].
b. *VP attachment*:
[$_{VP}$ Drag [$_{NP}$ the file] [$_{PP}$ next to the item]]].

While (1) is ambiguous with respect to syntax, to those with moderate, little, or no knowledge of the computer application being discussed, the sentence is also ambiguous with respect to meaning. Both the NP and the VP attachment of the PP could result in a viable interpretation of the sentence. But, because the two attachments have different meanings, if we wish to translate the sentence into a language that is not ambiguous in the same way English is, we must choose an attachment.

Japanese is just such a language, i.e., it is syntactically unambiguous with respect to PP attachment. Consider the Japanese sentences (2a/b), which correspond to the English sentences (1a/b), respectively.

(2a) (*NP-attachment case*)
［項目 の となり **の** ファイル］をドラッグ
item Gen next Gen file Acc drag
してください。
do-please
"(Please) drag the file that is next to the item."
(2b) (*VP-attachment case*)
［項目 の となり **に**］ファイル を ドラッグ
item Gen next to file Acc drag
してください。
do-please
"(Please) drag the file next to the item."

The difference between (2a) and (2b) is the postposition used with ファイル(*file*): in (2a), the postposition の(*of*) is used, whereas in (2b), the postposition に (*to*) is used. In Japanese, the presence or the absence of the postposition の plays a critical role in disambiguating PP attachment: the presence of の in (2a) indicates that the PP is attached to the object NP; the absence of の in (2b) indicates that the PP is attached to the VP. Since Japanese is unambiguous with respect to PP attachment, in order to produce a correct translation of an English sentence, we must disambiguate PP attachment in English.

2 Reattachment Strategy

Work for this paper was done in the context of the MT system at Microsoft Research (MSR-MT) (Menezes, A and S. Richardson, 2001). In this system, reattachment of English PPs takes place in the English analysis component after an initial parse is produced. By design, the initial parse has low right attachments of PPs. The reattachment module traverses the nodes of the parse tree and marks all the potential attachment sites for each PP. For instance, the only potential alternative attachment site for the PP in example (1) is that indicated in Figure 1 by a question mark.

Drag the file next to the item.

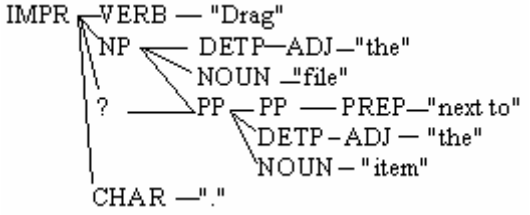

Figure 1: Alternative PP attachment site

After the reattachment module locates all the potential attachment sites, reattachment rules evaluate each site. These rules use syntactic information from the parse tree as well as punctuation and lexical information. One important piece of information that the rules get from the monolingual dictionary is the affinity of a noun/verb/adjective for a certain preposition. Below we give a snapshot of the relevant information in the verb entry for *remove*.

(3) Entry *remove*
Bitrecs {Bits T1 Hsubj
Vprp (from) }

This can be read as follows: *Remove* is used with *from* (Vprp) in contexts in which there is often a human subject (Hsubj) and an object (T1). Information of a similar type, though less detailed, is found on noun and adjective records as well.

Not surprisingly, our dictionary does not have all the verb-preposition information we need to make PP attachment decisions. To avoid the expensive and difficult hand-coding of the fine-grained lexical information that we need for PP attachment resolution, we followed the approach described in the next section.

3 Overview of Our Approach

Recently, a variety of approaches to the problem of PP attachment have been described in the literature. These fall into two categories, supervised and unsupervised. Among the unsupervised approaches, which use large, unanalyzed monolingual corpora, are those described in Hindle and Rooth (1993), Ratnaparkhi (1998), and Pantel and Lin (2000). Among the supervised approaches, which learn from disambiguated attachments, are those described in Stetina and Nagao (1997), Collins and Brooks (1995) and Brill and Resnik (1994).

Our approach, unlike those above, makes use of bilingual corpora as data. The approach is unsupervised, but it does require a large, parsed, sentence-aligned, bilingual corpus. We exploit the unambiguous nature of PP attachment in Japanese in our approach.

Just as our training data differs from those used in the systems mentioned above, so does our goal. Our aim is not to create a new algorithm for reattachment, but to collect information of the type used currently by our system in making reattachment decisions. Since our system favors low right attachments, our goal is to collect more information on when a verb has greater affinity for a PP than a noun does.

We used English-Japanese aligned parsed corpora, consisting of about 1 million sentences from computer manuals to extract two different types of data: (i) data that serve as positive evidence for VP attachment (ii) data that serve as negative evidence for VP attachment. Positive evidence consists of examples for which VP attachment is suggested by the Japanese data. Negative evidence consists of examples for which NP attachment is suggested.

3.1 Methodology

We began by parsing all the sentence pairs in the sentence-aligned bilingual corpus into high-level linguistic representations. These representations, called logical forms (LFs), represent the predicate argument structure of the input (Heidorn, G. E. 2000). By way of example, we give below the LF for the sentence in Figure 1:

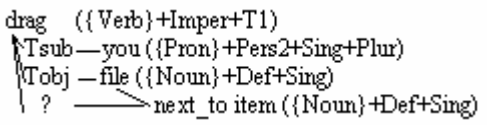

Figure 2: LF of "Drag the file next to the item."

The Japanese side of our bilingual corpus was parsed, with an attempt to identify the most likely attachment site for each prepositional phrase. The English side, however, was parsed with no attempt to find the correct attachment site of prepositional phrases. Instead, the lowest attachment site and other possible attachment sites, as determined by the first phase of reattachment, were computed, and these attachment sites were identified in the LF as shown by the question mark in Figure 2.

At this point, a modified version of the training phase of our example-based MT system was run over the LF pairs that are known to be mutual translations. As described in (Menezes, A and S. Richardson, 2001), this phase normally takes the aligned LF pairs, identifies correspondences between LF nodes using a combination of lexical and structural information, and then divides the aligned LF into smaller subsets that are to be used as translation mappings. Here, however, we exploited the existing training component only to identify the node correspondences in each aligned LF pair.

For each annotated pair, we found node subsets that fulfill the following conditions:
(i) There is an English NP (E_{N2}) that is the object of a prepositional phrase headed by the preposition *p*;
(ii) E_{N2} has a parent noun, E_{N1}, and an alternative attachment site, E_V, which is a VP;
(iii) E_{N2} is uniquely aligned to a Japanese node, J_1; and
(iv) J_1 has one parent, J_2, which is uniquely aligned to either E_{N1} or E_V.
From each such subset of nodes, we extracted a 5-tuple (E_V, E_{N1}, *p*, E_{N2}, V/N-attach), in which V/N-attach is set to V if J_1 aligns with E_V, and it is set to N if J_1 aligns with E_{N1}. In this way, we let the Japanese LF provide the answer as to whether an English prepositional phrase with preposition *p* is attached to a verb *v* or to an intervening noun n_1.

We then aggregated these statistics to compute a simple probability of a prepositional phrase headed by *p* attaching to a verb *v* in the following manner:
P(*p* attaches to *v*) = c(*v*, *, *p*, *, V) / c(*v*, *, *p*, *, *)
(where 'c' stands for count and '*' stands for wildcards)

3.2 Examples

To demonstrate the algorithm, we give two pairs of sentences, one from which we derive positive evidence for VP attachment, and the other from which we derive negative evidence for VP attachment. We obtained positive evidence for VP attachment from sentences (4a) and (4b):

(4) a. Type accidant in the document.
b. 文書に accidant と入力します。

The LFs for (4/ab), with arrows indicating how they are aligned, are given as Figure 3:

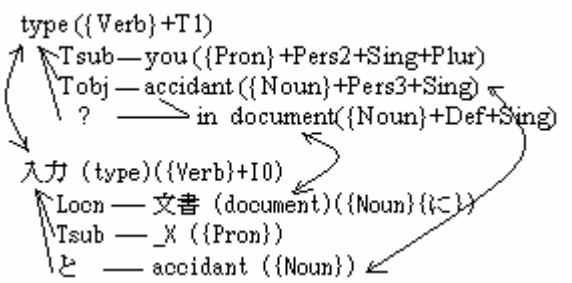

Figure 3: LF-alignment

In Figure 3, the English NP, *document* (E_{N2}), serves as the object of the preposition *in*; it has the parent NP, *accidant* (E_{N1}) and the alternative attachment site (E_V), which is marked by a question mark. E_{N2}, *document*, is aligned with the Japanese NP, 文書(J_1), and its parent (J_2), the verb, 入力, is aligned to the English verb *type*. This qualifies, therefore, as support for VP attachment, and the tuple (*type*, *accidant*, *in*, *document*, V-attach) is added to the training data.

From examples (5a/b) below, we obtained evidence for NP attachment. The aligned LFs for this pair appear as Figure 4.

(5) a. To access public data from a parent form.
b. 親フォームのパブリックデータに
アクセスするには

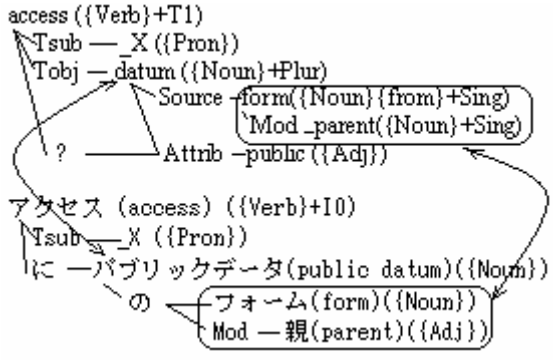

Figure 4：LF-alignment

The English NP *parent form* (E_{N2}) is the object of the preposition *from;* it has the parent NP, *public data* (E_{N1}) and the alternative attachment site, *access* (E_V), which is marked by a question mark. E_{N2}, *parent form*, is aligned with the Japanese NP, 親フォーム(J_1), and its parent NP (J_2), パブリックデータ, is aligned with the English NP attachment site, *public data*. This qualifies, therefore, as support for NP attachment, and the tuple (*access*, *datum*, *from*, *form*, N-attach) is added to the training data. Appendix A provides examples of VP/NP-attachment tuples collected in this way.

4 Incorporating the New Information into the System

Based on the algorithm described in Section 3.1, we extracted both positive evidence for verb-PP attachment (i.e., (v, n1, p, n2, V-attach)) and negative evidence (i.e., (v, n1, p, n2, N-attach)). From this we compiled a list of common verb-preposition pairs that are not in our monolingual dictionary. We excluded those verb-preposition pairs for which there were fewer than 10 instances of positive evidence. We also excluded those pairs for which the probability of VP attachment (as described in Section 3.1) was less than 50%. With these thresholds, we extracted 294 verb-preposition pairs (see Appendix B).[1] We created an auxiliary English dictionary with an entry for each verb in the 294 pairs. Each entry was populated, as in (3) above, with the preposition information from the verb-preposition list. The new dictionary was used in the reattachment module described in Section 2. A new reattachment rule assigned the highest possible score to a VP attachment site if the preposition of the PP considered for reattachment was in the list of prepositions associated with the verb in the auxiliary dictionary.[2]

5 Evaluation

We evaluated our new verb-preposition data and its use in reattachment of English PPs by evaluating its effect on English-Japanese (E-J) translation and English-Spanish (E-S) translation produced by MSR-MT system. Although evaluation of MT is a difficult problem, we chose MT as a vehicle for evaluating reattachment for two reasons. First, we are fortunate that, as a result of much attention given to the evaluation of MSR-MT translations, these evaluations are now carried out regularly in an efficient and effective way (Richardson, S. et. al. 2001, Pinkham, J and M. Corston-Oliver 2001). Second, the evaluation of PP attachment in its own right is extremely difficult for a native speaker of English, especially if the data is technical and the evaluator is not a technical expert.

Our evaluation was conducted as follows:

(i) We built two sets of the E-J translation database and two sets of the E-S translation database. One set was built using the English auxiliary dictionary containing new verb-preposition information. The other was built without this dictionary.

(ii) We ran through our translation systems enough English text from our blind corpus of technical texts to obtain 250 sentences that were translated differently by the two systems.[3]

(iii) A group of evaluators who are native or near-native speakers of the target languages were given the 250 English source sentences, a reference translation of each source sentence, and the two translations (presented in random order) of each source sentence.[4] They were asked both to judge

[1] We informally experimented with positive probability > 65% and positive probability > 50%. We found that with a threshold of 65%, too many valid verb-preposition pairs dropped from the list.

[2] We opted for this aggressive reattachment strategy after experimenting with other strategies in which we allowed for more interplay between the new and existing reattachment rules.

[3] For the E-J evaluation, we needed 2688 sentences to obtain 250 different translations and for the E-S evaluation, we needed 5344 sentences.

[4] For this evaluation, we are grateful to Mike Carlson, Mo Corston-Oliver and the people of the Butler Hill

which translation was closer to the reference translation (i.e., assign relative scores) and to rank the quality of both translations on a scale of 1-4 (i.e., assign absolute scores).[5]

Table 1 and Table 2 below show the results of the E-J and E-S evaluations, respectively. [6]

E-J Results

	Relative Scores (2688 sentences)	
With the auxiliary dict vs. Without the auxiliary dict	0.09033 (+/- 0.078) (with the auxiliary dict preferred = 1; without the auxiliary dict preferred = -1)	
	Absolute Scores (250 sentences)	
	Mean	Variance
With the auxiliary dict	2.580 (+/- 0.217)	0.590
Without the auxiliary dict	2.526 (+/- 0.231)	0.537

Table 1

E-S Results

	Relative Scores (5344 sentences)	
With the auxiliary dict vs. Without the auxiliary dict	-0.124 (+/- 0.103) (with the auxiliary dict preferred = 1; without the auxiliary dict preferred = -1)	
	Absolute Scores (250 sentences)	
	Mean	Variance
With the auxiliary dict	2.857 (+/- 0.239)	0.215
Without the auxiliary dict	2.89 (+/- 0.228)	0.234

Table 2

E-J translations produced by the system that used the auxiliary dictionary were judged to be significantly better than those produced from the system that did not use this dictionary, according to the relative score ($p = 0.012$). We did not see a significant difference in the absolute scores between the two systems.

The significant improvement with the auxiliary dictionary was in line with our expectations. The English analyses produced by our system originally had too many low, incorrect attachments. With an increased number of higher, correct attachments, English structures became more similar to Japanese structures. For a translation system that relies heavily on aligning linguistic structures to create translation mappings, getting linguistic structures to converge can help produce better translation mappings.

In the E-S evaluation, the system without the auxiliary dictionary was significantly better than the one with the dictionary, though again, the absolute scores showed no significant difference. This result was not unexpected. The Spanish analyses in our system have more low right attachments than our English analyses do. By producing more high PP attachments with the new auxiliary dictionary for English, we actually caused our Spanish and English linguistic structures to diverge. In an example-based MT system such as ours in which the training phase depends on both lexical information and structural correspondences, structural differences can hinder alignment and lessen the yield of the training phase.

Even if the divergence of linguistic structures were to have no negative effect on alignment, a better attachment in English would not guarantee a better translation into Spanish because it often does not matter if a PP is attached to a noun or to a verb; its translation into Spanish is the same.[7]

We do not claim, however, that correct English attachment is undesirable for our E-S translations. There are many cases in which correct attachment can make a positive difference in translation, as in example (6).

(6a) Source Sentence:
You can **open** the search **page from** the Web toolbar

Group.

[5] The absolute scores are as follows : 1 = unacceptable ; 2 = possibly acceptable ; 3 = acceptable ; 4 = ideal.

[6] The relative scores reflect the average of all raters on all sentences. Rather than evaluate sentences which were identical for both systems, we added enough "dummy" lines, with a relative score of 0 – that is, neither system preferred – to account for the sentences which were identical in the full sample size. The absolute scores, of course, only reflect the 250 sentences which were different in the two conditions.

[7] For example, (ia) is translated as (ib) whether *in a specific message* (en mensaje específico) is attached to *insert* (insertar) or to *signature* (firma).
(i) a. You can **insert** a **signature in** a specific **message**, or automatically add a signature to the end of every message.
(i) b. Puede **insertar** una **firma en** un **mensaje** específico o puede agregar una firma automáticamente al final de cada mensaje.

(6b) Reference Translation:
La **página** de búsqueda puede **abrirse desde** la barra de herramientas Web.
(the page of search can open from the bar of tools Web)
(6c) Translation using auxiliary dictionary:
La **página** de búsqueda **se puede abrir desde** la barra de herramienta Web.
(6d) Translation without auxiliary dictionary:
Puede abrir la página de búsqueda **de** la barra de herramienta Web

In (6), attachment makes a difference for the translation of the preposition. The preposition *from* in *open…from* is translated as *desde*, whereas *from* in *page from* is translated as *de.* An additional benefit of the correct attachment is that we get the passive reflexive *se* in the Spanish translation.

We might conclude from our E-J and E-S results that we should use MT to evaluate English reattachment only when one of the languages in the translation pair is unambiguous with respect to PP attachment, and we should not use MT to evaluate PP reattachment if both languages in the translation pair are ambiguous and they are disambiguated by the system to a different extent. A preferable conclusion, though, is that we need to use multilingual data to disambiguate PP attachment in Spanish as well as in English. Whereas there is something to say for maintaining ambiguity in translation, not all languages are ambiguous in the same way. Multilingual translations are going to require that we disambiguate all constructions in all languages to the greatest extent possible.

6 Future Work

As mentioned in the previous section, one task for the future is to disambiguate PP attachment in Spanish (and all other languages we work with that are ambiguous with respect to PP attachment) using data from the languages we work with (like Japanese) that are not ambiguous with respect to PP attachment.

An additional item for the future, which we have already begun, is to use bilingual parsed data to disambiguate other ambiguous constructions. We are currently working on disambiguating *–ing* forms in English. To exemplify this problem, we consider possible analyses of *switching equipment*: (i) *equipment* is the subject of *switch*; (ii) *equipment* is the object of *switch*; and (iii) *switching* is an underspecified verbal modifier of *equipment* (i.e., equipment for switching). For this problem we are fortunate to have not only Japanese data, but French and Spanish as well. None of these languages is characterized by the English ambiguity.

Finally, there is work to be done on PP attachment using the data we extracted from our aligned parsed database. With a focus on extracting useful data and working with, rather than drastically modifying, our current reattachment module, we made minimal use of the very specific lexical information we extracted. As can be seen in Appendix A and Appendix B, we originally extracted (v, n1, p, n2, V/N-attach) 5-tuples. While we used the frequency of VP vs. NP attachments to extract our verb-preposition pairs, there is no reason we cannot make full use of the n1 and n2 information in our 5-tuples.[8]

References

Brill, E. and Resnik, P. 1994. A Rule-based Prepositional Phrase Attachment Disambiguation. In *Proceedings 94*, Kyoto, Japan.

Collins, M. and Brooks, J. 1995. Prepositional Phrase Attachment through a Backed-off Model. In *Proceedings of the Third Workshop on Very Large Corpora*, pp. 27-38. Cambridge, Massachusetts.

Heidorn, G. E. 2000. Intelligence Writing Assistance. In Dale R., Moisl H., and Somers H. (eds.), A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text. Marcel Dekker, New York, 1998 (published in August 2000), pages 181-207.

Hindle, D and Rooth, M. 1993. Structural Ambiguity and Lexical Relations, *Computational Linguistics* 19(1): 103-120.

Menezes, A. and Richardson, S. 2001. A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. To appear in Proceedings of the ACL 2001, Toulouse, France.

[8] For instance, given the extracted 5-tuples of (v, n1, p, n2, attach) such as in Appendix A and with some similarity information of the type used by Pantel and Lin (2000), we could have prevented the VP attachment of the PP in (i) while allowing it in (ii):
(i) View [$_{NP}$ data in files].
(ii) View data [$_{pp}$ in browser].

Pantel, P and Lin, D. 2000. An Unsupervised Approach to Prepositional Phrase Attachment using Contextually Similar Words. In *Proceedings of Association for Computational Linguistics 2000*: pp. 101-108, Hong Kong.

Pinkham, J and M. Corston-Oliver. 2001. Adding Domain Specificity to an MT system. Proceedings of the Workshop on Data-Driven Machine Translation, ACL 2001, Toulouse, France.

Ratnaparkhi, A. 1998. Unsupervised Statistical Models for Prepositional Phrase Attachment. In *Proceedings of COLING-ACL 98*. Montreal, Canada.

Richardson, S. & Dolan, W. & Menezes, A. & Corston-Oliver, M. 2001. Overcoming the Customization Bottleneck Using Example-based MT. In Proceedings of the Workshop on Data-Driven Machine Translation, ACL Conference, June 2001.

Steina, J. and Nagao, M. 1997. Corpus Based PP Attachment Ambiguity Resolution with a Semantic Dictionary. In *Proceedings of the Fifth Workshop on Very Large Corpora*: 66-80, Beijing and Hong Kong.

Appendix A:

Partial list of VP/NP Attachment Data

verb	n1	p	n2	V/N-attach
access	class	from	scope	V
access	computer	from	workgroup	V
access	information	from	database	N
access	namespace	from	service	N
add	entry	for	service	V
add	example	for	syntax	V
add	entry	for	datatype	N
add	file	for	project	N
configure	computer	with	setting	V
configure	protocol	with	address	V
configure	pool	with	range	N
configure	use	with	connection	N
specify	time	in	second	V
specify	switch	in	command	V
specify	setting	in	file	N
specify	tab	in	box	N
update	permission	in	FrontPage	V
update	style	in	manner	V
update	table	in	dataset	N
update	text	in	frame	N

Appendix B:

Partial list of verb-preposition pairs in the auxiliary dictionary

access from	extract from	reset to
access through	import from	resolve to
add as	import into	restore from
add for	import to	restore to
add in	include in	restrict to
add on	insert in	retrieve from
add to	insert into	return as
append to	install as	return for
apply to	install for	return in
assign to	install from	return with
attach to	install in	route to
change in	install into	run against
change on	install on	run for
change to	install to	run in
check against	limit to	run on
choose from	link to	save as
choose in	list in	save in
click for	load from	save to
click in	load into	search in
click on	log as	see for
click under	log in	select as
click with	log to	select for
configure as	manage from	select from
configure for	map to	select in
configure in	merge into	select on
configure on	migrate to	select under
configure with	modify as	send to
connect to	modify in	separate with
contact for	move to	set as
contain for	obtain from	set for
contain in	open from	set in
control from	open in	set on
convert into	open on	set to
convert to	paste into	specify as
copy from	perform for	specify for
copy into	perform in	specify from
copy to	perform on	specify in
create for	place in	specify on
create from	place on	start from
create in	print on	start in
create on	protect from	start on
create with	provide for	store as
define as	provide in	store in
define for	publish in	store on
define on	publish to	support for
delete from	read from	support on
designate as	receive from	treat as

determine on	record in	type as
disconnect from	redirect to	type at
display as	refer for	type from
display in	reference from	type in