Locations in the Machine Translation of Prepositional Phrases

Arturo Trujillo

Computer Laboratory
University of Cambridge
Cambridge CB2 3QG, England
iat@cl.cam.ac.uk

Abstract

An approach to the machine translation of locative prepositional phrases (PP) is presented. The technique has been implemented for use in an experimental transfer-based, multi-lingual machine translation system. Previous approaches to this problem are described and they are compared to the solution presented.

1 Introduction

In the construction of a multi-lingual machine translation (MT) system it is important to maintain the independence of the monolingual components. The main reason for this is that grammar components developed in a monolingual context can be added with relatively few modifications to the same MT system.

Two ways of achieving this are 1) using an Interlingua as a common representation, and 2) having independent bilingual and monolingual modules in the system. An example of the former is the system described in [Farwell and Wilks, 199l] where sentences in all languages are converted into a language-independent formalism. Such a representation is then passed to any other language component for generation into the target language (TL). Other examples of this type of solution include Principle-Based MT [Dorr, 1990] and Knowledge-Based MT [Mitamura *et al.,* 1991].

The second type of solution is also exemplified in more than one way. For example, in a transfer based system such as Eurotra [Arnold and des Tombe, 1987], [Maegaard and Perschke, 1991], the system developed at ISSCO [Russell *et al.,* 1991] and the system developed by SRI International [Alshawi *et al.,* 1991], there are two types of rules: those which depend on one language and those which depend on two languages. The former are normally syntactic and semantic rules and the latter are usually called transfer rules. As another example, in the type of statistical MT described in [Brown *et al., 1990],* there are two statistical models: the language model which contains only monolingual information and the translation model which contains purely bilingual information.

The system described in this paper is a transfer-based system where the only source of bilingual information is the bilingual lexicon, and where individual monolingual components are constructed almost independently of each other (they are not totally independent because there must be agreement upon the structure of the information used at the point of transfer). The system follows the lexicalist approach to MT suggested in [Tsujii and Fujita, 1991] and [Beaven, 1992] where transfer rules reduce to bilingual lexicon equivalences. In addition, dependencies between words in a sentence are expressed by indices as in [Zeevat, 1991].

The bilingual lexicon is to resemble human bilingual dictionaries as much as possible. The reason to favour a lexically-driven approach is that bilingual dictionaries are becoming available in machine readable form and it would be an advantage to use them as sources of bilingual knowledge. In addition, the entries in the bilingual lexicon will be reversible, in accordance with the arguments put forward in [Isabelle, 1989].

2 Locative PP's

Locative PP's are those which are used to specify the physical location of an action or an object. The following are some examples taken from [Sparck-Jones and Boguraev, 1987]:

Mr Brown is at the office
She sat by the pillar
Sebastian felt pain in his foot

The type of situations each locative PP can describe varies widely from context to context; this is well demonstrated in [Herskovits, 1986]. in this paper little will be said about the way context influences the translation of a locative PP.

3 Approaches to PP Translation

There are two points to consider in PP translation: representation and disambiguation. By representation it is meant the formal structures used for representing PP's. Disambiguation concerns the information needed for selecting appropriate target language (TL) prepositions. The ideal position is one where the representation in the source language (SL) is completely independent of the TL, and where there is enough information present in the representation for selecting the correct translation. Note that we are not considering PP attachment at the monolingual level since this may be best carried out independently of TL, and there have been several proposals for handling it including [Wilks *et* a/., 1985] using preference semantics, [Jensen and Binot, 1987] using semantic information in on-line dictionary definitions and [Whittemore *et* a/., 1990] using an experimentally determined combination of heuristics containing lexical, syntactic, linear and semantic information.
This paper is concerned mainly with the representational issue.

3.1 Interlingua PP Translation

The representation of PP's in Interlingua systems involves either deep cases and universal predicates, or ideal meanings and conceptualizations. An explanation of each now follows.

Universal Predicate PP Translation

Deep cases and universal predicates are used as language independent relations representing the meaning of a preposition. For example, in [Barnett *et* a/., 1991] the preposition *on* in the sentence 'there were five fish on the plate' is represented by the universal predicate *supported-by z x y,* where *z, x* and *y* are the discourse markers for the state, the fish and the plate respectively.

Some deep cases used widely include *locative, instrumental* and *temporal.* For a discussion on cases see [Somers, 1987]. When there is a one-to-many correspondence between the Interlingua representation and the TL, disambiguation is done using semantic information. For example, in [Jin, 1991] the semantic class (eg. force, abstract property, abstract object) of arguments in the Interlingua representation is used for selecting TL words; this presumably includes the selection of prepositions.

Ideal Meanings PP Translation

Ideal meanings and conceptualizations is the approach adopted in [Japkowicz and Wiebe, 1991] based on [Herskovits, 1986] and [Grimaud, 1988]. Here an objective representation of the situation described by the PP is constructed using the ¡deal meaning of the preposition and the possible language dependent conceptualizations of its complement noun phrase (NP). For instance, the objective representation of the sentence *she is on the bus* would include among other things that a bus has a platform and a volume, and that the subject of the sentence is contained in that volume. Given this language-independent representation, a TL preposition is selected as follows: from the TL lexicon, the conceptual representation of the complement NP is retrieved; this would include a highlighted feature which in English would be the platform of the bus and in French it would be its volume. Then, a TL preposition that can have such a conceptual representation as its complement is selected. This preposition must also be appropriate for the spatial situation being described.

Consider, for example, the translation of the above sentence from English into French. The French lexicon would state that *bus* can be conceptualized as a volume, and since the objective or real-world

representation of the situation states that *she is* contained by the bus, the French preposition *dans* is selected. Another approach along the same lines but used in a transfer system with semantic features instead of the semantic network used by [Japkowicz and Wiebe, 1991] is presented in [Zelinsky-Wibbelt, 1990].

Since generation in this type of solution occurs from an objective or real-world meaning representation, all the information needed for correct TL selection is present in the representation. The ambiguities that remain are ambiguities present in the SL: if the TL allows similar ambiguities then they are preserved across translation, otherwise, multiple translations are produced or one is selected by default or using a heuristic. For instance, in the previous example the French equivalent of the reading where *she* is on top of the bus would also be generated but later discarded.

3.2 Transfer Based PP Translation

In a transfer system such as Eurotra [Durand *et a*/., 1991], PP's which are arguments (in the Eurotra sense of argument, *Ibid.* p. 113) to a word have their prepositions recoded into the feature bundle of the Interface Structure (IS) representation of the word (e.g. *depend on* becomes *{lu=depend, pform-of-arg2=on}).* Adjunct PP's are encoded with the preposition as governor of the IS representation of the NP. TL selection in t(ranslation)-rules is achieved at the IS level by using semantic features from the NP and from the phrase to which the PP attaches.

4 Suggested Approach

The implemented system makes use of transfer rules stated as translation equivalences relative to SL and TL lexical entries. This distinguishes it from Interlingua based systems by not postulating universal predicates, and from standard transfer-based systems by not recoding structures during analysis of the SL and by disallowing tree-to-tree transformations. An example and description of how the approach works now follows. Take as input the following sentence:

Eng: Mary rests by the pillar

1) Analysis produces a representation which is independent of the TL.

$\{mary_1 , rest_{2,1} , by_{2,3,4} , the_4 , pillar_4\}$

where the { } represent a bag. The indices represent dependency relations between the words in the input sentence.

2) The bilingual lexicon has the following equivalences:

$\{mary_x\} \Leftrightarrow \{maría_x\}$

$\{rest_{x,y}\} \Leftrightarrow \{descansar_{x,y}\}$

$\{by_{x,y,z}\} \Leftrightarrow \{a_{x,y} , el_y , lado_y , de_{y,x}\}$

$\{the_x\} \Leftrightarrow \{el_x\}$

$\{pillar_x\} \Leftrightarrow \{pilar_x\}$

Using these translation equivalences, the transfer stage produces the following representation:

Spa: $\{maria_1 , descansar_{2,1} , a_{2,3} , el_3 , lado_3 , de_{3,4} , el_4 , pilar_4\}$

3) This representation is used as input to a generator (in effect a modified parser of functionality equivalent to the one presented in [Reape, forthcoming]) which tries to construct a sentence from a bag of lexemes by arranging them into an order which is licensed by the grammar. After morphological generation (which is not done in the current version of the system) the output is:

María descansa al lado del pilar

Lit. Mary rests to-the side of-the pillar

4.1 Comment

The idea behind the assignment of indices in 1) is that within a sentence there are a number of dependency relations which are to be made explicit by the monolingual grammar. The way the system actually produces the indices is by parsing the input using the type system and parser of the LKB [Copestake, 1992] and then assigning unique identifiers (cf. Skolemising, number_var in Prolog) to each shared distinct index. For example, the simplified rule below, when given as input the NP *the cat,* constructs the representation {the$_x$, cat$_x$}. This representation is then passed to a function which assigns a unique identifier to each distinct shared index to give {the$_1$, cat$_1$}.

NP[rep: <d1> U <d2>] ⇒
  Det[rep: <d1>={det$_x$}]
  Noun[rep: <d2>={noun$_x$}]
  (angle brackets represent a shared structure).

The indices used belong to three sorts: 'temporal', 'object' and 'location'. 'Temporal' indices will correspond roughly to what are sometimes called 'events'; they are used to index verbs and adverbs. 'Object' indices are used for nouns, determiners and adjectives. It is worth mentioning that at the moment these sorts are only used as a guide to assigning indices to lexical items. In [Zeevat, 1991] they are used to capture certain semantic entailments; here they simply serve to indicate dependency relations within a phrase.

4.2 Locations in PP MT

In order to cope with PP's an additional index sort, namely 'location', is introduced. Its use is exemplified as the index *y* in the transfer rule:

{by $_{x,y,z}$} ⇔ {a$_{x,y}$, el$_y$, lado$_y$, de$_{y,z}$}

Locations are regions in space associated with an object. They are normally expressed by prepositions, although they may appear in nouns, as in Spanish *lado* in the rule above. Note that it is possible that an index of sort 'location' can not be motivated in Spanish for the noun *lado;* in this case it would be necessary to preserve the index number but not its sort during translation.

As argument fillers, locations have been suggested as a way of analysing locative PP's in English: in [Sondheimer, 1978, p. 246] they are used as the first argument to a preposition predicate (e.g. in *the park* becomes *IN(p,the park)* where 'p' is a place referent); their existence is supported by certain referring expressions such as *here, there, everywhere, somewhere,* etc., [Jackendoff, 1983, pp. 50-55]; they have been important to certain semantic theories, such as Situation Semantics, in their analysis of locative PP's (e.g. Colban's analysis in [Fenstad *et al.,* 1987, Appendix A]); and they have been used in implemented systems for expressing semantic entailments from sentences containing locative PP's, [Creary *et al.,* 1989].

The usefulness of locations in MT is twofold. First they allow TL independence. For example, consider the translation of the preposition *by* presented above, which does not have a one word translation into Spanish in its locative sense. It is tempting to introduce in the Spanish grammar a rule which will construct a single predicate which is equivalent in meaning to the English preposition *by,* say *al-lado-de.* This is fine for purely bilingual translation, but note that now we have made the Spanish grammar contain English (i.e. TL) information, namely the fact that English has one word to express the multiple word phrase *al lado de.* Now, when we try to add a new language to the system we may encounter the following situations: a constructed predicate is not necessary for translating to or from the new language, or a new constructed predicate needs to be introduced into the existing grammars to cope with a single predicate in the new language, or both. An example of the first case would occur if we added Portuguese to the system, where the translation of Spanish *al lado de* is the literal translation *ao lado do;* now, an unnecessary ambiguity between Portuguese and Spanish translation could arise, since the system may translate the Spanish phrase either literally or using the constructed predicate needed for English-Spanish translation. An example of the second problem would be the incorporation of Hungarian into the system. In Hungarian the single word *honnan* can be translated into English as

from where and into Spanish as *de donde.* In a system which allows TL knowledge in the monolingual grammars, two new predicates, say *from-where* and *de-donde* would have to be added to the English and Spanish grammars respectively. In reality both situations are due to the same problem, namely introducing predicates in one language which are motivated by another language.

The other advantage of using locations is that they can be used to represent prepositions that allow PP's instead of NP's as their complements. Within an MT context this type of constructions has been noticed by [Durand *et* a/., 1991, p. 119], from where the following examples are taken:

Out from under the bed
Researchers from within the community (ESPRIT corpus)

The way this is aided by locations is that we can represent the complex preposition in the second example compositionally as:

$\{ \text{from}_{1,2,3} , \text{within}_{2,3,4} \}$

where 2 and 3 are locations, 1 would be bound to *researchers* and 4 to *community.* Now the following transfer rules would suffice for transferring this complex preposition to give the translation below.

$\{\text{from}_{x,y,z}\} \Leftrightarrow \{\text{de}_{x,y,z}\}$
$\{\text{within}_{x,y,z}\} \Leftrightarrow \{\text{dentro}_{x,y}, \text{de}_{y,z}\}$

Spa. Investigadores de dentro de la comunidad
Lit. Investigators of inside of the community

Note the two instances of *de:* as case marker with 2 indices, and as locative preposition indicating source, with 3 indices.

5 Comparison with Other Approaches

As described in the previous section, the two main advantages of the approach to PP translation described in this paper are: 1) it allows a language independent treatment of the translation of prepositions which do not have a corresponding preposition in the TL (e.g. *by),* and 2) it provides a way for handling compound prepositions (e.g. *from within).* To contrast this with the approaches described earlier let us see how they might handle these phenomena.

In the Universal Predicate or Deep Case approach, prepositions are mapped to language independent predicates, usually expressing binary relations. To correctly translate a preposition which does not have a corresponding preposition in the TL, it will be necessary for the TL grammar to contain a rule mapping the language independent predicate into the TL. For instance, if we assume a predicate *proximate x y* for the sense of *by* above, there will have to be a rule in the Spanish grammar of the form:

[... sem: proximate(x,y) ...] -> [orth: al lado de ...]

Clearly, adding a new language to the system would require adding rules of this form to the monolingual grammars; this in effect loses some of the independence of monolingual components. Compound prepositions, unless a new predicate is used for each possible combination, may be represented in one of three ways; for *from within* in the example above these might be:

1) source(researchers,community) & contained-by-boundary(researchers,community)
2) source(researchers,contained-by-boundary(researchers,community))
3) source(researchers,contained-by-boundary(community))

The problem with 1) is that it states that the researchers are within the community. 2) states that the researchers are from researchers within the community, which is slightly unsatisfactory since it introduces an extra group of researchers. The solution in 3) requires every universal predicate to have two entries, one for when it occurs in a compound preposition, and one for when it does not.

Turning now to the Ideal Meanings approach, the main difference here is that the actual translation of prepositions makes no use of a bilingual lexicon but of language independent meanings. It is argued here that, if 'locations' are used, PP translation can be carried out using the bilingual lexicon only. The main advantage of this is that use can be made very readily of human bilingual lexicons in the construction of multilingual MT systems.

As far as structural Transfer approaches are concerned, the problem of prepositions without a direct translation can be tackled by equating the preposition in question with a structure of equivalent meaning in the other language. For instance, in a system that effects transfer at the level of syntactic trees, the following kind of transfer rule may be necessary (Ox represents a translation variable):

(PP by @x) ↔ (PP de (IMP (Det el) (N (AP otro) (N (N lado) (PP de @x)))))

The main disadvantage of this strategy is that these rules are usually very unrestricted and could give rise to non-terminating computations. When transfer is effected at higher levels of representation, such problems may be avoided, but then the problems begin to resemble those given for Deep Case approaches.

6 Problems

In the current version of the system, the transfer and generation algorithms are both quite inefficient in the worst case. In the case of generation, the inefficiency stems from generating by rearranging a bag of words into a grammatical sequence. The condition for termination, then, is not only that the sequence be licensed by the grammar, but also that all the lexemes in the input to the generator be consumed. For instance, given the bag

$\{the_1, red_1, car_1, stop_{2,1}\}$

the system constructs the two sentences

the red car stops
the car stops

This is because both are licensed by the grammar and it is not known at the point of constructing the subject of the sentence whether the NP contains an adjective or not. Hence, the lower sentence must be discarded after generation on the grounds that the number of predicates consumed by the whole sentence is not the same as the number in the input bag.

Another problem with the generation algorithm is that it makes it difficult to insert lexical items which are not predictable from lexical-transfer information alone. For example in the translation

Eng: the big fat cat sleeps
Spa: el gato grande y gordo duerme
Lit: the cat big and fat sleeps

there is no way of predicting the behaviour of the conjunction *y* using lexical information alone. For example, translating into Spanish, we will get a bag containing the above Spanish words except *y;* consequently, the system will fail to generate the appropriate sentence. A solution to this might be to allow the insertion of *y* in the Spanish bag when translation is from English (this might be incorrect for, say, Italian-Spanish translation, and should therefore be language pair specific).

The biggest source of inefficiency during transfer is the large number of possible translations that a SL predicate can have. This is a linguistic problem which can not be overcome by using a more efficient algorithm. It can only be handled by using semantic and pragmatic information to select the correct translation. However, there is no space to treat TL disambiguation here at length. What can be mentioned is that this type of information can be incorporated into the framework. This may be implemented as a different module which would be superimposed on the bilingual lexicon and used for disambiguation.

7 Conclusion

A representation for the translation of locative PP's has been presented which uses locations as a way of maintaining the independence of monolingual components and the compositionality of the representation. Examples of why both are desirable were given.

The reversibility of the system is achieved by restricting transfer to substitution of lexical items using a bilingual lexicon. The bilingual lexicon was closely modelled on human bilingual dictionaries to ease their incorporation when large machine-readable versions become available. One consequence of this is that translation between similar languages (e.g. Spanish and Portuguese) is aided if consideration is given only to the information needed for transfer between these two languages.

Future research includes identifying the information necessary for TL selection of locative prepositions and investigating the possibility of using the indexing technique for other types of PP including temporal and causative ones.

Acknowledgements

This work was funded by the UK Science and Engineering Research Council. Many thanks to Ted Briscoe, Antonio Sanfilippo, John Beaven, Ann Copestake, Valeria de Paiva, and three anonymous reviewers. Thanks also to Trinity Hall, Cambridge, for a travel grant. All remaining errors are mine.

References

[Alshawi *et al.,* 1991] H. Alshawi, D. Carter, B. Gamback, and M. Rayner. Swedish-English QLF translation. In H. Alshawi, editor, *The Core Language Engine,* chapter 14. MIT Press, Cambridge, MA, USA, 1991.

[Arnold and des Tombe, 1987] D. Arnold and L. des Tombe. Basic theory and methodology in Eurotra. In S. Nirenberg, editor, *Machine Translation, Theoretical and Methodological Issues,* chapter 7. Cambridge University Press, Cambridge, England, 1987.

[Barnett *et al.,* 1991] J. Barnett, I. Mani, E. Rich, C. Aone, K. Knight, and J. C. Martinez. Capturing language-specific semantic distinctions in interlingua-based MT. In *Proceedings MT Summit III,* Washington DC, July 1991.

[Beaven, 1992] J. L. Beaven. Shake-and-bake machine translation. In *COLING '92,* Nantes, France, July 1992.

[Brown *et al.,* 1990] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational Linguistics,* 1990. Vol. 16(2) pp. 79-85.

[Copestake, 1992] A. Copestake. The AQUILEX LKB: Representation issues in semi-automatic aquisition of large lexicons. In *Proceedings 3rd Conference on Applied Natural Language Processing,* Trento, Italy, March-April 1992.

[Creary *et al.,* 1989] L. G. Creary, J. M. Gawron, and J. Nerbonne. Reference to locations. In *Proceedings ACL-89,* Vancouver, Canada, June 1989.

[Dorr, 1990] B. J. Dorr. Machine translation: A principle-based approach. In P. H. Winston and S. A. Shellard, editors, *Artificial Intelligence at MIT, Expanding Frontiers, Vol. 1,* chapter 13. The MIT Press, Cambridge, Mass., USA, 1990.

[Durand *et al.,* 1991] J. Durand, P. Bennett, V. Allegranza, F. van Eynde, L. Humphreys, P. Schmidt, and E. Steiner. The Eurotra linguistic specifications: An overview. *Machine Translation,* 1991. Vol. 6, No. 2 pp. 103-147.

[Farwell and Wilks, 199l] D. Farwell and Y. Wilks. Ultra: A multi-lingual machine translator. In *Proceedings MT Summit III,* Washington DC, July 1991.

[Fenstad *et al.,* 1987] J. E. Fenstad, P. Halvorsen, T. Langholm, and J. van Benthem. *Situations, Language and Logic.* D. Reidel Publishing Co., Dordrecht, Holland, 1987.

[Grimaud, 1988] M. Grimaud. Toponyms, prepositions and cognitive maps in English and French. *Journal of the American Society of Geolinguistics,* 1988.

[Herskovits, 1986] A. Herskovits. *Language and Spatial Cognition: An Interdisciplinary Study of the Prepositions in English.* Studies in Natural Language Processing. Cambridge University Press, Cambridge, England, 1986.

[Isabelle, 1989] P. Isabelle. Toward reversible MT systems. In *Proceedings MT Summit II,* Munich, Germany, 1989.

[Jackendoff, 1983] R. Jackendoff. *Semantics and Cognition.* The MIT Press, Cambridge, MA, 1983.

[Japkowicz and Wiebe, 1991] N. Japkowicz and J. M. Wiebe. A system for translating locative prepositions from English into French. In *Proceedings of the ACL 91,* Berkeley, CA, June 1991.

[Jensen and Binot, 1987] K. Jensen and J. Binot. Disambiguating prepositional phrase attachments by using on-line dictionary definitions. *Computational Linguistics,* 1987. Vol. 13, No. 3-4 pp. 251-260.

[Jin, 1991] W. Jin. Translation accuracy and translation efficiency. In *Proceedings MT Summit HI,* Washington DC, July 1991.

[Maegaard and Perschke, 1991] B. Maegaard and S. Perschke. Eurotra: General system design. *Machine Translation,* 1991. Vol. 6(2) pp. 73-82.

[Mitamura *et al.,* 199l] T. Mitamura, 3rd E. H. Nyberg, and J. G. Carbonell. An efficient interlingua translation system for multi-lingual document production. In *Proceedings MT Summit III,* Washington DC, July 1991.

[Reape, forthcoming] M. Reape. Parsing semi-free word order and bounded discontinuous constituency and "shake 'n' bake" machine translation (or 'generation as parsing'). In M. Emele, U. Heid, S. Momma, and R. Zajac, editors, *Proceedings of the Workshop on Constraint-based Approaches to Natural Language Generation,* Bad Teinach, Germany, forthcoming.

[Russell *et al.,* 1991] G. Russell, A. Ballim, D. Estival, and S. Warwick. A language for the statement of binary relations over feature structures. In *Proceedings of the European Chapter of the ACL,* Bonn, Germany, April 1991.

[Somers, 1987] H. Somers. *Valency and Case in Computational Linguistics.* Edinburgh University Press, Edinburgh, Scotland, 1987.

[Sondheimer, 1978] N. K. Sondheimer. A semantic analysis of reference to spatial properties. *Linguistics and Philosophy,* 1978. Vol. 2(2) pp. 235-280.

[Sparck-Jones and Boguraev, 1987] K. Sparck-Jones and B. Boguraev. A note on a study of cases. *Computational Linguistics,* 1987. Vol. 13(1-2) pp. 65-68.

[Tsujii and Fujita, 1991] J. Tsujii and K. Fujita. Lexical transfer based on bilingual signs: Towards interaction during transfer. In *Proceedings European ACL-91,* Berlin, Germany, August 1991.

[Whittemore *et al.,* 1990] G. Whittemore, K. Ferrara, and H. Brunner. Empirical study of predictive powers of simple attachment schemes for post-modifier prepositional phrases. In *28th Annual Meeting of the ACL,* Pittsburgh, PA, June 1990.

[Wilks *et al.,* 1985] Y. Wilks, X. Huang, and D. Fass. Syntax, preference and right attachment. In *IJCAI-85,* Los Angeles, CA, 1985.

[Zeevat, 1991] H. W. Zeevat. *Aspects of Discourse Semantics and Unification Grammar.* PhD thesis, University of Amsterdam, Amsterdam, The Netherlands, 1991.

[Zelinsky-Wibbelt, 1990] C. Zelinsky-Wibbelt. The semantic representation of spatial configurations: a conceptual motivation for generation in machine translation. In *COL1NG '90,* Helsinki, Finland, August 1990.