Spatial Lexicalization in the Translation of Prepositional Phrases

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

Abstract

A pattern in the translation of locative prepositional phrases between English and Spanish is presented. A way of exploiting this pattern is proposed in the context of a multilingual machine translation system under development.

Introduction

Two of the main problems in machine translation (MT) are ambiguity and lexical gaps. Ambiguity occurs when a word in the source language (SL) has more that one translation into the target language (TL). Lexical gaps occur when a word in one language can not be translated directly into another language. This latter problem is viewed by some as the key translation problem, (Kameyama *et al.*, 1991).

A case in point is the translation of prepositional phrases (PP). The following entry for the translations into Spanish of the preposition *along* demonstrates this (entry taken from (Garcia-Pelayo, 1988)).

along: por (by), a lo largo de (to the length of), según (according to)

Both problems occur here: there are three different translations for the same English preposition, and the second of these is a phrase used to describe a sense of *along* which is not encoded as one word in Spanish.

Lexicalization Patterns

It is argued in (Talmy, 1985) that languages differ in the type of information they systematically encode in lexical units. That is, languages exhibit distinct lexicalization patterns. For instance, in a sentence where both the direction and manner of motion are expressed, English will encode motion and manner in the same verb, whereas in Spanish a distinct lexicalization of these two meaning components will be favoured (*Ibid.* p. 69):

Spa. El globo **subió** por la chimenea flotando
Lit. the balloon **moved-up** through the chimney floating
Eng. The balloon floated up the chimney

*This work was funded by the UK Science and Engineering Research Council

Here Spanish *subió* encodes 'move + up' whereas English *floated* encodes 'move + floating'.

Capturing lexicalization patterns of this sort can help us make certain generalizations about lexical gaps and ambiguities in MT. In the rest of this paper two lexicalization patterns for English locative prepositional phrases (PP) will be presented. It will be shown how they allow us to simplify the bilingual lexicon of a transfer based, multi-lingual MT system under development.

Evidence

The two lexicalization patterns under analysis can be illustrated using the following three sentences (loc = location, dest = destination):

Eng. She ran **under**$_{loc}$ the bridge (in circles)
Spa. Corrió **debajo del** puente (en círculos)
Lit. Ran-she under of-the bridge

Eng. She ran **under**$_{path+loc}$ the bridge (to the other side)
Spa. Corrió **por debajo del** puente (hasta el otro lado)
Lit. Ran-she along under of-the bridge

Eng. She ran **under**$_{dest+loc}$ the bridge (and stopped there)
Spa. Corrió **hasta debajo del** puente (y allí se detuvo)
Lit. Ran-she to under of-the bridge

In the first sentence there is a direct translation of the English sentence. In this case the features encoded by the English and Spanish PP's are the same. In the second sentence the English preposition encodes the path followed by the runner and the location of this path with respect to the bridge; in Spanish such a combination needs to be expressed by the two prepositions *por* and *debajo de*. In the third example the English preposition expresses the destination of the running and the location of that destination with respect to the bridge; this has to be expressed by the two Spanish prepositions *hasta* and *debajo de*.

Other English prepositions which allow either two or three of these readings in locative expressions are shown in the table below.

P	location	path 'along P'	destination 'to P'
behind	detrás de	por detrás de	hasta detrás de
below	debajo de	por debajo de	hasta debajo de
inside	dentro de	por dentro de	hasta dentro de
outside	fuera de	por fuera de	hasta fuera de
under	debajo de	por debajo de	hasta debajo de
between	entre	por entre	-
near	cerca de	-	hasta cerca de

From the table the following generalization can be made: whatever the translation *P* of the locative sense of an English preposition is, its path incorporating sense is translated as *por P* and its destination incorporating sense is translated as *hasta P*.

In short, certain English prepositions are ambiguous between encoding location, path + location or destination + location. This is not the case in Spanish. When translating from English such ambiguities can not be preserved very naturally. In particular, whenever it is necessary to preserve them (e.g. for legal documents), a disjunction of each individual sense must be used in the TL sentence.

In certain cases, however, it may be the case that only one of these readings is allowed.

Disambiguation

As far as the selection of the appropriate target language (TL) preposition is concerned the constituent which the PP modifies plays a major role in determining which readings of a preposition sense are allowed.

Deciding whether the preposition is used in a spatial sense, as opposed to a temporal or causative sense, is determined by the semantics of the noun phrase (NP) within it, e.g. *under the table, under the regime, under three minutes, under pressure, under development, under the bridge*; that is, a place denoting NP gives rise to a spatial PP.

There are two cases to consider in disambiguating spatial senses. In the case of the PP attaching to a noun, the sense selected will be the location one. For example

Eng. The park **outside** the city

Spa. El parque **fuera de** la ciudad

The second case is when the PP modifies a verb. For this case it is necessary to consider the semantics of the verb in question. Verbs of motion such as *walk, crawl, run, swim, row, gallop, march, fly, drive, jump* and *climb* allow location, path and destination readings. For instance:

Eng. The diver swam **below** the boat

Spa. El buceador nadó **debajo de/por debajo de/hasta debajo de/l** bote

Verbs which do not express motion such as *stand, sit, rest, sleep, live* and *study* usually require the location sense of the preposition:

Eng. The diver rested **below** the boat

Spa. El buceador descansó **debajo del** bote

This second analysis is oversimplistic since some readings depend on other semantic features of the verb, preposition and complement NP involved. However, these can be incorporated into the strategy explained below.

One last point to note is that not all the prepositions presented allow all three readings. This will be taken into consideration when making the generalizations in the encoding of the above observation.

Encoding

Representation for Prepositions

As exemplified above, the translation of a preposition depends on three sources of information: 1) the word modified by the PP determines whether the sense of the preposition may include a path or a destination component, 2) the preposition itself determines how many spatial senses it allows, 3) the NP complement of the preposition determines whether it is being used spatially, temporally, causatively, etc. To encode these three sources, prepositions will be represented as three place relations. The pattern for a prepositional entry is shown in 1); a possible entry for *below* is shown in 2).

1) *P[modified,preposition,complement]*
2) below[motion-verb,[path,dest],place]

The notation here is an informal representation of the typed feature structures described in (Briscoe *et al.*, 1992) and (Copestake, 1992). The argument types in 1) can be explained as follows. 'Modified' is a type which subsumes 'events' (denoted by verbs) and 'objects' (denoted by nouns); the type 'event' is further subdivided into 'motion-verb' and 'non-motion-verb'. 'Preposition' is a type which subsumes properties which depend on the preposition itself; for the examples presented this type will encode whether the preposition can express a path or a destination (the extra square brackets indicate a complex type). Finally, 'complement' subsumes a number of types corresponding to the semantic field of the complement NP; these include 'spatial' with subtype 'place'; 'temporal', and 'causative'.

The instantiated entry in 2) corresponds to the use of *below* in *the diver swam below the boat.* Such instantiations would be made by the grammar by structure sharing of the semantic features from the modified constituent and from the complement NP. In this way the three translations of *below* would only be produced when the semantic features of the modified constituent and complement NP unify with the first and third arguments respectively.

Bilingual Lexical Rules

To encode the regularity of the translations presented, bilingual lexical rules will be introduced. These rules take as input a bilingual lexical entry and give as output a bilingual lexical entry. An oversimplified rule to generate the 'path' sense for a preposition that allows such a reading is given below (P = variable ranging over prepositions, e = the empty type, *lugar* = place, *camino* = path).

Rule:

P_{Eng}[motion-verb,[path,-],place] ↔
P_{Spa}[verbo-movimiento,e,lugar] de

⇓

P_{Eng}[motion-verb,[path,-],place] ↔
POR[verbo-movimiento,camino,lugar]
P_{Spa}[verbo-movimiento,e,lugar] de

A similar rule would encode the 'destination' sense generalization.

The bilingual lexical rules work by extending the bilingual lexicon automatically before any translation takes place; this gives rise to a static transfer component with faster performance but more memory consumption. Only those entries which unify with the input part of a rule actually produce a new bilingual entry.

An example of the 'path' rule being applied is shown below.

Input:

below[motion-verb,[path,dest],place] ↔
debajo[verbo-movimiento,e,lugar] de

Output:

below[motion-verb,[path,dest],place] ↔
POR[verbo-movimiento,camino,lugar] debajo[verbo-movimiento,e,lugar] de

Note that not all prepositions in the table above allow all three readings; for this the allowed readings are stated in the second argument of the preposition.

Related Research

In (Copestake *et al.*, 1992) the notion of a *tlink* is introduced. These are typed feature structures which encode generalizations about the type of transfer relations that occur in the bilingual lexicon. That is, each bilingual entry corresponds to one *tlink*. Because *tlinks* are represented as a hierarchy of types, the amount of data stored in the bilingual lexicon is minimal. The bilingual lexical rules presented here will further refine the idea of a *tlink* by minimizing the number of bilingual lexical entries that have to be coded manually, since the bilingual lexical rules can be seen as operating over *tlinks* (and hence bilingual lexical entries) to give new *tlinks*.

The grammatical formalism used broadly resembles earlier versions of HPSG. The idea of bilingual lexical rules is partly inspired by the lexical rules introduced within this framework in (Pollard & Sag, 1992).

Conclusion

We have argued that ambiguities and lexical mismatches found in English-Spanish translation of PP's can be dealt with using ideas from cross-linguistic studies of lexicalization patterns, and suggested a use of the relevant linguistic insights for MT applications.

This consisted of encoding prepositions as three place relations, and of having bilingual lexical rules which operate over the bilingual lexicon to expand it. By formulating regularities in this way consistency and compactness in the bilingual lexicon, and therefore in the transfer module, are achieved.

The next steps will include the implementation of the mechanism to drive the bilingual lexical rules, the refining and testing of the semantic classification, the isolation of further regularities and the investigation of other types of PP's.

Acknowledgements

Many thanks to Ted Briscoe, Antonio Sanfilippo, Ann Copestake and two anonymous reviewers. Thanks also to Trinity Hall, Cambridge, for a travel grant. All remaining errors are mine.

References

Briscoe, T.; Copestake, A., and de Paiva, V., editors. 1992 (forthcoming). *Default Inheritance in Unification Based Approaches to the Lexicon.* Cambridge University Press, Cambridge, England.

Copestake, A.; Jones, B.; Sanfilippo, A.; Rodriguez, H.; Vossen, P.; Montemagni, S., and Marinai, E. 1992. Multilingual lexical representations. Technical Report 043, ESPRIT BRA-3030 AQUILEX Working Paper, Commission of the European Communities, Brussels.

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

Garcia-Pelayo, R. 1988. *Larousse Gran Diccionario Español-Inglés English-Spanish.* Larousse, Mexico DF, Mexico.

Kameyama, M.; Ochitani, R., and Peters, S. 1991. Resolving translation mismatches with information flow. In *Proceedings ACL-91*, Berkeley, CA.

Pollard, C., and Sag, I. 1992 forthcoming. *Agreement, Binding and Control: Information Based Syntax and Semantics Vol. II.* Lecture Notes. CSLI, Stanford, CA, USA.

Talmy, L. 1985. Lexicalization patterns: semantic structure in lexical forms. In Shopen, T., editor, *Language Typology and Syntactic Description Vol. III: Grammatical Categories and the Lexicon.* Cambridge University Press, Cambridge, England.