Inflectional Information in Transfer for Lexicalist Machine Translation

DAVIDE TURCATO, OLIVIER LAURENS,
PAUL MCFETRIDGE & FRED POPOWICH
Simon Fraser University, Canada

Abstract

We show how an MT system can perform the transfer of structurally divergent inflectional information in a simple and modular way, while placing no constraint on the format in which such information is expressed on either monolingual side. We take advantage of the flexibility of the Shake-and-Bake approach and extend the expressive capabilities of bilingual entries in several directions.

1 Introduction

The transfer of inflectional information such as verb tense and aspect is important in order to obtain successful and natural translations. Problems underlying the analysis of inflection include mismatches and structural divergences across languages. Furthermore, because inflectional information is often separately handled in pre-parsing and post-generation stages, the format in which it is expressed is often unrelated to and less predictable than the format of syntactic information.

As Van Eynde (1995:67-68) points out, traditional transfer approaches achieve an adequate treatment of inflectional structural divergences either by having very complex transfer components or by adopting multistratal monolingual components, which derive intermediate abstract representations. Such representations are less divergent, and thus more manageable, than their respective surface structures. This is the approach taken, for instance, in Eurotra (Van Eynde 1988) and in Verbmobil (Dorna & Emele 1996).

In the more recent literature on transfer-based MT, attempts have been made to preserve simplicity in the transfer component by using monostratal grammars. Most notably:

1. Van Eynde (1995) proposes a sign-based approach in which transfer of tense and aspect information is achieved in a declarative fashion by means of value sharing between features in the source and target verb representations.

2. The Shake-and-Bake (henceforth S&B) approach to MT (Beaven 1992, Whitelock 1994) adopts a variation of the transfer paradigm where only lexical transfer rules are used — there are no structural transfer rules. These lexical transfer rules can contain explicit inflectional information and even independent morphemes. The lexical transfer rules (also known as bilingual lexical entries) establish relationships between source language and target languages lexemes: one or more source language lexemes are mapped to one or more target language lexemes. Specifically, a bag (multiset) of lexemes created during the analysis of the source language sentence or phrase, together with the bilingual lexical entries, are used to produce a bag of target language lexemes. This operation is performed by unifying the bilingual lexical entries lexemes (represented as feature structures) with those contained in the bag which then become 'consumed' (they are no longer available for transfer). The target language bag is then used as input (together with a target language grammar and lexicon) to a generator to create a target language sentence or phrase. Under this approach, inflectional information has no special status. It is regarded as lexical information and its transfer is performed in the same modular way in which syntactic information is transferred.

The drawback of both the approaches described above is that they have to place heavy restrictions on their monostratal grammars in order to achieve simplicity in transfer. More specifically:

1. Van Eynde (1995) relies on the assumption that the same theoretical background and formalism, namely HPSG, are used for both the source and target languages.
2. The standard S&B architecture assumes a concatenative morphology. In an S&B system with no capability to state generalizations over bilingual entries, this step is necessary in order to accommodate transfer of inflectional information between non-equivalent lexical items, i.e. source and target lexical items which are not paired in the same bilingual entry[1].

In this paper we will describe a framework for lexicalist transfer MT, which permits the combination of monostratal grammars and simple transfer procedures without having to place any constraints on the kind of grammars in use. The issue is particularly relevant if we consider real-world, large-scale, multilingual MT systems. In such systems, it is vital to be able to reuse

[1] An attempt to overcome such a restriction can be found in Turcato (1995).

existing linguistic resources rather than developing resources from scratch. This results in a system architecture which must take into account the features of pre-existing grammars, lexicons and inflectional processors. Our proposal will take the S&B approach as its starting point and extend it in several directions.

In the following, we will first examine the issues concerning the transfer of inflectional information. Then we will propose an approach to deal with these issues, which will be followed by some examples showing how translation problems involving tense and aspect can be elegantly handled.

2 Inflectional information

As an integral module of the syntactic component (Anderson 1982) the expression of inflectional morphology will vary with constraints on syntactic structure. In particular, we find:

1. divergences between the inflectional systems of different languages. In most cases there is no isomorphism between the two inflectional systems, however they are described. Some sort of information can be present on one side but not on the other, giving rise to gaps and mismatches. E.g., information about verb perfectivity is inflectionally absent from English but present in Spanish. Therefore, transfer between the two systems is a more complex correspondence than a simple one-to-one mapping;
2. divergences in the way the same inflectional information is realized in the sentence. For instance, the same sort of information can be realized as a separate word in one case and as inflection of a base form in the other (e.g., auxiliaries vs. suffixes);
3. divergences in the attachment of inflectional information. Structural divergences between languages often require that equivalent inflectional information be attached to non-equivalent stems (e.g., when head switching occurs). Instead of being regarded as units, inflected forms need to be broken into separate components;
4. concurrence between monolingual and bilingual constraints to determine inflectional realization. The information for proper inflectional realization comes only partially from transfer. It must be integrated with monolingual information during generation in order to perform a correct inflection.

The treatment of inflectional information changes according to different theories and different languages. In the case of English, for instance, its poor

inflectional system makes it possible to include inflected forms as separate lexical entries, thus avoiding any specific inflectional treatment. In the case of highly inflectional languages, like Spanish or Portuguese, such an approach would be very impractical. Such languages require inflectional generalizations which allow the derivation of inflected forms from base forms. Different approaches come in at this point: a major distinction is that between morpheme based approaches, where inflected forms are derived by concatenation of base forms and morphemes, and approaches where other devices (e.g., lexical rules) are used in place of morphemes. In the latter case, although inflected forms are derived from base forms, inflectional information never appears in the form of actual separate morphemes.

In the context of a large scale natural language translation system from English to a highly inflectional language like Spanish or Portuguese, the availability of various English dictionaries containing inflected word forms eliminates the need for inflectional analysis of the source language sentence — the inflectional information will be present in the feature structures associated with the English words. On the target side, an inflectional processor (morphological generator) can derive inflected forms from base forms and associated feature structures. The actual theoretical approach taken by the inflectional generators does not matter with respect to our discussion. For our present purposes it will suffice to say that the transfer module has only access to inflectional information in the form of features in lexical signs.

3 Lexicalist inflectional transfer

Our proposal will extend the standard S&B approach in the following directions:

1. We augment bilingual entries so that they may also contain *transfer attachments*. The attachments may place constraints on the source and target bags, or may add constituents to either side of the bilingual entry.
2. We augment the transfer module to allow the definition of user defined *transfer macros*. Transfer macros are essentially just parameterized versions of a traditional bilingual entry.

Transfer attachments are associated with bilingual entries by means of a 'double backslash' operator ('\\'). The attachments either consist of transfer macro calls (see Section 3.1) or of bag constraints (which will be discussed within Section 3.2).

The transfer of inflectional information is thus performed by transfer macros, rather than by independent morpheme-based bilingual entries, with macro calls appearing as transfer attachments in bilingual lexical entries. Thus, feature-based inflection systems can be accommodated in transfer. A successful treatment of inflection on these terms allows the accommodation of HPSG (Pollard & Sag 1994) and other formalisms which don't use a concatenative approach to inflection. In general, the approach allows us to deal with complex transfer cases, in which the transfer of information (not just inflectional information) involves non equivalent lexical items.

3.1 *Transfer macros*

The use of macros in grammars and lexicons is not new. A transfer macro is simply a parameterized bilingual entry. When a macro call appears on a transfer attachment associated with an ordinary bilingual entry, it can add lexical items on either side (or viewed procedurally, have the transfer rule consume additional lexical items on the source side and add lexical items to the target side) or perform an additional feature transfer for the lexical items passed to it as parameters. A key feature of a transfer macro is that it can contain a condition on the lexical items passed to it as parameters. In this way the transfer macros can apply conditionally, if the parameters satisfy the condition. Therefore, a transfer macro can comprise several clauses, each of which caters for a specific case. This is very useful, for instance, when transferring tense and aspect. The same transfer macro, say `trans_verb`, is associated with every bilingual entry involving verbs; the sign representing the verb is passed to the macro as a parameter. The appropriate clause is then used, depending on the actual content of the relevant features in the verbal parameters.

Transfer macros are the means by which we state generalizations over bilingual entries. In this respect, they play a similar role to that of bilingual lexical rules, as described in (Trujillo 1995).

3.2 *Translation of tense and aspect*

It will be shown here how the transfer of verb form information can be performed using a feature-based morphology, and we will illustrate the role of bag constraints. A detailed discussion will be provided for the English-Spanish language pair. A sample of relevant sentences will be provided and the transfer procedure for the sentences will be outlined. Our aim is not to give a complete account of the complex issue of temporal relations, but

rather to provide a range of examples aimed at showing that the framework described here can satisfactorily handle whatever information is needed for an adequate treatment of inflection.

The general form of a bilingual entry for a verb is the following (for the sake of simplicity we represent a one-to-one entry, but we assume that additional lexical items can be present on either side of the double arrow):

(1) `Eng_word::(Eng_desc)` $\leftrightarrow$ `Spa_word::(Spa_desc)`
`\\trans_verb(Eng_desc,Spa_desc).`

where `trans_verb` is a transfer macro call attached to the bilingual entry. A transfer macro definition has the following form:

(2)
$$
\begin{array}{l}
\texttt{macro\_name(Eng\_arg}_1\texttt{, Spa\_arg}_1\texttt{) tmacro} \\
\texttt{Eng\_word}_{1,1}\texttt{::(Eng\_desc}_{1,1}\texttt{)} \\
\texttt{\&} \ \ldots \\
\texttt{\& Eng\_word}_{1,j_1}\texttt{::(Eng\_desc}_{1,j_1}\texttt{)} \\
\leftrightarrow \\
\texttt{Spa\_word}_{1,1}\texttt{::(Spa\_desc}_{1,1}\texttt{)} \\
\texttt{\&} \ \ldots \\
\texttt{\& Spa\_word}_{1,k_1}\texttt{::(Spa\_desc}_{1,k_1}\texttt{)} \\
\vdots \\
\texttt{macro\_name(Eng\_arg}_n\texttt{, Spa\_arg}_n\texttt{) tmacro} \\
\texttt{Eng\_word}_{n,1}\texttt{::(Eng\_desc}_{n,1}\texttt{)} \\
\texttt{\&} \ \ldots \\
\texttt{\& Eng\_word}_{n,j_n}\texttt{::(Eng\_desc}_{n,j_n}\texttt{)} \\
\leftrightarrow \\
\texttt{Spa\_word}_{n,1}\texttt{::(Spa\_desc}_{n,1}\texttt{)} \\
\texttt{\&} \ \ldots \\
\texttt{\& Spa\_word}_{n,k_n}\texttt{::(Spa\_desc}_{n,k_n}\texttt{)}
\end{array}
$$

Any number of <word,description> pairs can appear to the left or right of the double arrow, even zero, as shown in example (5). Therefore either side can:

1. only set features on its argument;
2. only set one or more additional lexical items;
3. combine the two operations above.

Although this is not shown in the example above, a transfer macro can also contain transfer attachments and can thus call a further transfer macro. For instance, `trans_verb` calls the transfer macro `trans_tense`.

3.2.1 *Perfective vs. imperfective aspect*

The translation of the past tense from English to Spanish is an example of a mismatch between two inflectional systems.

(3) a. *He once spoke with me* ↔ *Habló conmigo una vez.*
b. *He spoke with me every day* ↔ *Hablaba conmigo cada día.*

(4) *He used to speak with me every day* ↔ *Hablaba conmigo cada día.*

The English past tense corresponds to two Spanish tenses, the preterite and the imperfect. Spanish adds to the verbs in the past tense the dimension of perfectivity, which is absent from English past tense verbs. However, an imperfective action can be signaled in English by the modal *used to*. In our examples, both English and Spanish tenses are typed features in verb signs. However, the value of the Spanish `tense` feature, instead of being atomic as on the English side, is a feature structure defined for boolean features like `past` and `perfective`[2]. Therefore, the type `preterite` is defined `[+past, +perfective]`, whereas `imperfect` is `[+past, -perfective]`. We take advantage of this distinction by transferring the English past tense to either a fully specified or an underspecified Spanish tense, depending on the information available on the English side.

In example (4), the specifications `[+past]` and `[-perfective]` (i.e. `[tense:imperfect]`) are set on *hablar*, depending on the presence of *used*, and *to*, coindexed with *speak* in some specified way, on the English side. This can be accomplished by a `trans_verb` clause like the following:

(5)
```
trans_verb((@index(X)), (tense:imperfect))
tmacro
used::@index(X) & to::@index(X) ↔ []
```

In cases like (3), the English past tense is transferred to an underspecified `[+past]` Spanish tense. During generation, the Spanish underspecified tense can get instantiated for its `perfective` feature, depending on additional information like adverbials or conjunctions, or either defaulted to some value (probably `[+perfective]`). It is worth noting that devices for the instantiation of the `perfective` feature can be implemented at any time on a purely monolingual ground, without affecting the transfer procedure.

3.2.2 *Auxiliaries vs. inflection*

The example below shows the alternation between separate lexical items and base form inflection in different languages, in order to express the same

[2] In the following, we adopt the notations +F and -F as shorthands to represent a feature F with a `plus` or `minus` value, respectively.

kind of information (the future tense, in this case).

(6) *He will speak tomorrow* ↔ *Hablará mañana.*

This pattern can be easily handled by a transfer macro which adds the lexical item *will* on the English side (coindexing it with the English verb argument) and a `[tense:fut]` feature on its Spanish argument, as follows:

(7)
```
trans_verb((@index(X)), (tense:fut))
tmacro
will::@index(X) ↔ []
```

A reverse situation is presented by the following example, where a simple English verb (*mending*) is mapped onto a compound Spanish verb:

(8) *My shoes need mending* ↔ *Mis zapatos necesitan ser reparados.*

In this case, the set of feature specifications which signal the 'passive present participle' on the English side are mapped onto a 'past participle' specification plus the addition of the auxiliary *ser* on the Spanish side, as follows:

(9)
```
trans_verb((@pass_pres_part), (@index(X),@past_part))
tmacro
[] ↔ ser::@index(X)
```

3.2.3 *Head switching*

The example below is traditionally shown as an example of head switching,

(10) *Mary swam across the river* ↔ *Mary cruzó el río nadando.*

The `[tense:past]` specification associated with *swim* needs to be transfered to a `[tense:+past]` specification on *cruzar*. However there is no bilingual entry which pairs *swim* and *cruzar*. Instead, *swim* is paired with *nadar* and *cruzar* with *across*.

In this case we resort to the capability of the bilingual entries to express constraints on source or target bags. This mechanism, which finds application in a larger range of phenomena than just inflectional transfer, resembles the use of *contextual variables* described by Trujillo (1995). In addition to expressing constraints on the actual lexical items paired in a bilingual entry, additional constraints can be expressed on lexical items to be found in the bags. Typically, such constraints involve sign descriptions, rather than words, but this is not a mandatory restriction.

We can formulate our account of bag constraints by saying that each side of a bilingual entry is simply a set of lexical items matching bag items. The only distinction is between items which are consumed from the bag (the

actual items of the bilingual entry) and items which are not consumed (the bag constraints). However, we find more appropriate a clearcut distinction between the two kinds of information for reasons of modularity. We will express bag constraints by means of a bag_cons attachment associated with a bilingual entry.

As to the head switching case described above, the inflectional transfer is triggered by the entry which pairs *across* and *cruzar* and is performed as follows: the tense on *cruzar* is transferred from a source bag item which satisfies the conditions of being a verb and being coindexed in some specified way with *across* (see variables A, B and C).

(11)
```
across::(E, @index(A,C)) ↔ cruzar::(S, @verb(A,B,C))
\\bag_cons(eng,(E,@verb(A,B))),
\\trans_verb(E,S).
```

Moreover, there is an entry that translates *swim* into *nadar* and that specifies the Spanish verb as 'gerund' if it is the translation of *swim* as in *swim across* (in other contexts, the tense and aspect of the English verb must be transferred the normal way). This is performed by the trans_verb transfer macro which states a bag constraint: the source bag must contain a directional adverbial such as *across*. This adverbial must, furthermore, be coindexed with the verb.

(12)
```
swim::(Eng, @verb(A,B)) ↔ nadar::(Spa, @verb(A,B))
\\trans_verb(Eng, Spa).
```

(13)
```
trans_verb(@index(A), (@index(A), @gerund))
tmacro
[] ↔ []
\\bag_cons(eng, (@directional_pp(A))).
```

This definition of trans_verb sets the form of the Spanish verb to 'gerund' if the English bag contains a directional adverbial (other clauses in the definition of trans_verb transfer the tense and aspect from the Englishverb otherwise). The above entries assume the same indices on *swim*, *across*, *cruzar* and *nadar*, under the assumption that *across* is a modifier of *swim* and *nadar* is a modifier of *cruzar*.

4 Conclusions

The formal devices we have described here do not imply any assumptions about what inflectional information should look like. The same machinery can equally handle feature-based and morpheme-based morphology, because

transfer macros can equally perform feature transfer or add extra lexical items on either side. Even a feature-to-morpheme or morpheme-to-feature transfer would be equally feasible. Actually, the pattern implicit in our examples implies both morphemes and features on both sides, since inflectional information like verb aspect can be represented by auxiliaries or modals on one side, but not on the other. Moreover, the format of the inflectional information and the way it is transferred are transparent to the bilingual entries to which the transfer macros are associated. Different treatments of inflection could be implemented without affecting the bilingual lexicon for base forms. What would change is only the content of the transfer macros.

While increasing the expressive power of a lexicalist MT system, the proposed approach does not increase its complexity. With respect to the standard S&B architecture, the only substantial addition is that of bag constraints. Their use can be expected to reduce the number of candidate target bags, filtering out those which do not satisfy the constraints. As to the other devices described here, they do not affect the performance of a system. Although a bilingual lexicon is described as a structured object, where transfer attachments can be embedded into one another, it is worth pointing out that a flat bilingual lexicon can still be compiled from one as described here. The extra level of structure allows information to be packaged in a more compact way, avoiding redundancy, but the computational workload at runtime is not heavier than in an equivalent lexicalist system with no transfer attachments.

Likewise, the reversibility and declarativity of a S&B system are not affected by the introduction of transfer attachments. A bilingual entry still states a relation between two bags, regardless of the direction and process at hand. The only difference between including a lexical item in the body of a bilingual entry or in a transfer attachment is that in the latter case the lexical item does not 'consume' a bag item. In other words, the lexical item must match a bag item consumed by some other bilingual entry independently triggered. Apart from marking such distinction, transfer attachments have no other purpose than allowing a more efficient information packaging, as pointed out above. Most fundamentally, the basic difference between S&B and traditional transfer systems is retained: transfer is a mapping between bags, not trees. No structural transfer is performed, hence no recursive traversal of structural representations is needed.

A transfer module developed according to the proposed guidelines can fit a very large range of monolingual grammars, lexicons and morphological processors, since no formal or theoretical assumption is made as to what

the monolingual components should look like. Empirical support to our claim is provided by the fact that the described guidelines have been implemented in a multilingual MT system, described in Popowich et al. (1997), in which pre-existing lexicons and morphological processors have been used and grammars have been independently developed according to different theoretical approaches.

Acknowledgements. This research was supported by a Collaborative Research and Development Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC), and by TCC Communications.

REFERENCES

Anderson, Stephen R. 1982. "Where Is Morphology". *Linguistic Inquiry* 13.571-612.

Beaven, John L. 1992. "Shake and Bake Machine Translation". *Proceedings of the 14th International Conference on Computational Linguistics (COLING-92)*, ed. by Christian Boitet, 603-609. Nantes, France.

Dorna, Michael & Martin Emele. 1996. "Semantic-based Transfer". *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, 316-321. Copenhagen, Denmark.

Pollard, Carl & Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar.* (= *Studies in Contemporary Linguistics*). Chicago, Illinois, U.S.A.: The University of Chicago Press.

Popowich, Fred, Davide Turcato, Olivier Laurens, Paul McFetridge, J. Devlan Nicholson, Patrick McGivern, Maricela Corzo-Pena, Lisa Pidruchney & Scott MacDonald. 1997. "A Lexicalist Approach to the Translation of Colloquial Text". *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'97)*, 76-86. Santa Fe, New Mexico, U.S.A.

Trujillo, Arturo. 1995. "Bi-Lexical Rules for Multi-Lexeme Translation in Lexicalist MT". *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'95)*, 48-66. Leuven, Belgium.

Turcato, Davide. 1995. "Shake-and-Bake MT and Morphology". *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'95)*, 319-325. Leuven, Belgium.

Van Eynde, Frank. 1988. "The Analysis of Tense and Aspect in Eurotra". *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)* ed. by Dénes Vargha, 699-704. Budapest, Hungary.

_____. 1995. "A Sign-Based Approach to the Translation of Temporal Expressions". *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'95)*, 67-86. Leuven, Belgium.

Whitelock, Pete. 1994. "Shake and Bake Translation". *Constraints, Language and Computation* ed. by C.J. Rupp, M.A. Rosner & R.L. Johnson, 339-359, London, U.K.: Academic Press.