IL MT System. Evaluation for Spanish-English Pronominal Anaphora Generation*

Jesús Peral and Antonio Ferrández

Departamento de Lenguajes y Sistemas Informáticos
Universidad de Alicante. Alicante, Spain
{jperal, antonio}@dlsi.ua.es

Abstract. In this paper the pronominal anaphora generation module of a complete interlingua Machine Translation (MT) approach is presented. The approach named AGIR (*Anaphora Generation with an Interlingua Representation*) allows the generation of anaphoric expressions into the target language from the interlingua representation of the source text. AGIR uses different kinds of knowledge (lexical, syntactic, morphological and semantic information) to solve the Natural Language Processing (NLP) problems of the source text. The paper presents the evaluation of the generation of English and Spanish (including zero pronouns) third person personal pronouns into the target language. The following results have been obtained: a precision of 80.39% and 84.77% in the generation of Spanish and English pronominal anaphora respectively.

1 Introduction

The problem of anaphoric expressions is one of the most difficult to solve in Natural Language Processing (NLP). This problem must be treated as two different processes: resolution and generation. The first one searches for the discourse entity to which the anaphor refers to. On the other hand, the process of generation consists on the creation of references over a discourse entity.

In the context of Machine Translation (MT) the resolution of anaphoric expressions is of crucial importance in order to translate/generate them correctly into the target language. After evaluating many commercial and experimental MT systems we observe that one of the main problems of them is that they do not carry out a correct pronominal anaphora generation. Solving the anaphora and extracting the antecedent are key issues in a correct generation into the target language. Unfortunately, the majority of MT systems do not deal with anaphora resolution and their successful operation usually does not go beyond the sentence level. This paper presents a complete approach that allows pronoun resolution and generation into the target language.

AGIR (*Anaphora Generation with an Interlingua Representation*) system works on unrestricted texts unlike other systems, the KANT interlingua system [10], the Météo system [3], the Candide system [2], etc. that are designed

* This paper has been partially supported by the Spanish Government (CICYT) project number TIC2000-0664-C02-02.

for well-defined domains. Although full parsing of these texts could be applied, we have used partial parsing of the texts due to the unavoidable incompleteness of the grammar. This is a main difference with the majority of the interlingua systems such as the DLT system based on a modification of Esperanto [17], the Rosetta system which is experimenting with Montague semantics as the basis for an interlingua [1], the KANT system, etc. as they use full parsing of the text.

After the parsing and solving pronominal anaphora, an interlingua representation of the whole text is obtained. From this interlingua representation, the generation of anaphora (including intersentential anaphora), the detection of coreference chains of the whole text and the generation of Spanish zero-pronouns into English have been carried out, issues that are hardly considered by other systems. Furthermore, this approach can be used for other different applications, e.g. Information Retrieval, Summarization, etc.

In the following section (section 2), the complete approach that includes Analysis and Generation modules will be described. These modules will be explained in detail in the next two sections. In section 5, the Generation module has been evaluated in order to measure the efficiency of our proposal. Finally, the conclusions of this work will be presented.

2 AGIR Architecture

AGIR system architecture (figure 1) is based on the general architecture of a MT system which uses an interlingua strategy. Translation is carried out in two stages: from the source language to the interlingua, and from the interlingua into the target language. Modules for analysis are independent from modules for generation. In this paper, although we have only studied the Spanish and English languages, our approach is easily extended to other languages, i.e. multilingual system, in the sense that any analysis module can be linked to any generation module. As can be observed in figure 1, there are two independent modules in the process of generation: Analysis and Generation modules.

3 AGIR's Analysis Module

In AGIR, the analysis is carried out by means of SUPAR (*Slot Unification Parser for Anaphora resolution*) system [5]. SUPAR is a computational system focused on anaphora resolution. It can deal with several kinds of anaphora, such as pronominal anaphora, one-anaphora, surface-count anaphora and definite descriptions. In this paper, we focus on pronominal anaphora resolution and generation into the target language. The input of SUPAR is a grammar defined by means of the grammatical formalism SUG (*Slot Unification Grammar*). A translator that transforms SUG rules into Prolog clauses has been developed. This translator will provide a Prolog program that will parse each sentence. SUPAR allows to carry out either a full or a partial parsing of the text, with the same parser and grammar. Here, partial parsing techniques have been used due to the

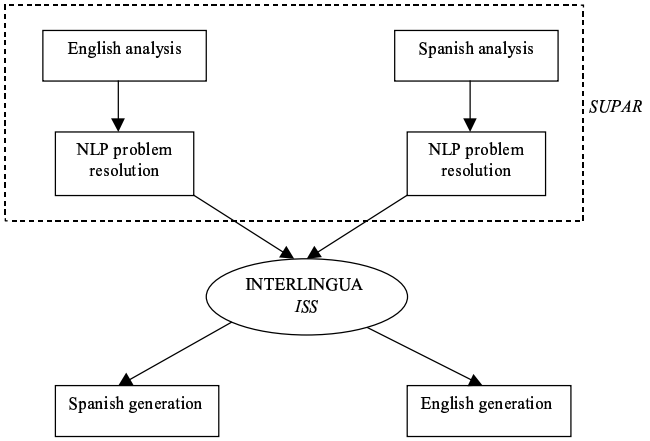

Fig. 1. AGIR architecture

unavoidable incompleteness of the grammar and the use of unrestricted texts (corpora) as inputs.

The analysis of the source text is carried out in several stages. The first stage of the analysis module is the lexical and morphological analysis of the input text. Due to the use of unrestricted texts as input, the system obtains the lexical and morphological information of the text's lexical units from the output of a part-of-speech (POS) tagger. The word, as it appears in the corpus, its lemma and its POS tag (with morphological information) is supplied for each lexical unit in the corpus.

The next stage is the parsing of the text (it includes the lexical and morphological information extracted in the previous stage). The corpus is split into sentences before applying the parsing. The output will be the Slot Structure (SS) that stores the necessary information[1] for NLP problem resolution.

In the third stage a module of Word Sense Disambiguation (WSD) is used to obtain a single sense for the different text's lexical units. The lexical resources WordNet [9] and EurowordNet [16] have been used in this stage.

The SS enriched with the information of previous stages will be the input for the following stage in which NLP problems (anaphora, extraposition, ellipsis, etc.) will be treated and solved.

After the anaphora resolution stage, a new Slot Structure (SS') is obtained. In this new structure the correct antecedent (chosen from the possible candidates after applying a method based on restrictions and preferences [5]) for each anaphoric expression will be stored together with its morphological and semantic

[1] The SS stores for each constituent the following information: constituent name (NP, PP, etc.), semantic and morphological information, discourse marker (identifier of the entity or discourse object) and the SS of its subconstituents.

information. The new structure SS' will be the input for the last stage of the Analysis module.

In the last stage AGIR generates the interlingua representation of the whole text. This is the main difference between AGIR and the rest of MT systems that carry out a processing of the input text sentence by sentence. The interlingua representation will allow the correct generation of the intrasentential and intersentential pronominal anaphora into the target language. Moreover, AGIR allows the identification of coreference chains of the text and their subsequent generation into the target language.

The interlingua representation of the input text is based on the clause as main unit of this representation. Once the text has been split into clauses, AGIR uses a complex feature structure for each clause. It is composed of semantic roles and features extracted from the SS of the clause. Semantic roles that have been used in this approach are the following: ACTION, AGENT, THEME and MODIFIER that correspond to verb, subject, object and prepositional phrases of the clause respectively. The notation we have used is based on the representation used in KANT interlingua. To identify these semantic roles when partial parsing has been carried out and no semantic knowledge is used, the following heuristic has been applied:

***H*$_1$** Let us assume that the NP parsed before the verb is the agent of the clause. In the same way, the NP parsed after the verb is the theme of the clause. Finally, all the PP found in the clause are its modifiers.

It is important to emphasize that the interlingua lexical unit has been represented in AGIR using the word and its correct sense in WordNet. After accessing to the ILI (Inter-Lingual-Index) module of EuroWordNet, we will be able to generate the lexical unit into the target language.

Once the semantic roles have been identified, the interlingua representation will store the clauses with their features, the different entities that have appeared in the text and the relations between them (such as anaphoric relations). This representation will be the input for the generation module. More details about the interlingua representation in AGIR have been presented in [15, 13].

4 AGIR's Generation Module

The interlingua representation of the source text is taken as input of the Generation module. The output of this module is the target text, that is, the representation of the source text's meaning with words of the target language. In this paper we are only describing the generation of third person personal pronouns into the target language, so we have only focused on the differences (syntactic and morphological) between the Spanish and English languages in the generation of the pronoun. These differences are what we have named discrepancies (a detailed study of Spanish-English-Spanish discrepancies is shown in [15, 13]).

4.1 Syntactic Discrepancies

Elliptical Zero-subject Constructions (Zero-pronouns) The Spanish language allows to omit the pronominal subject of the sentences. These omitted pronouns are usually named zero pronouns. While in other languages, zero pronouns may appear in either the subject's or the object's grammatical position, (e.g. Japanese), in Spanish texts, zero pronouns only appear in the position of the subject. In [14, 6] the processing of Spanish zero pronouns in AGIR is presented. Basically, in order to generate Spanish zero pronouns into English, they must first be located in the text (ellipsis detection), and then resolved (anaphora resolution). At the ellipsis detection stage, information about the zero pronoun (e.g. person, gender, and number) must first be obtained from the verb of the clause and then used to identify the antecedent of the pronoun (resolution stage).

Pleonastic Pronouns Sometimes pronouns can be used in a non-referential construction, that is, appear due to some requirement in the grammar of the language. These pronouns are usually named pleonastic. In AGIR, the pleonastic use of pronoun *it* has been detected before the anaphora resolution stage and thereby will not be resolved. These pronouns will appear marked like pleonastics in the interlingua representation, they will not have antecedent and they will not be generated into Spanish. In order to detect pleonastic *it* pronouns in AGIR, a set of rules, based on pattern recognition, that allows the identification of this kind of pronouns is constructed. These rules are based on the study developed by other authors [8, 11] that faced with this problem in a similar way.

4.2 Morphological Discrepancies

Number Discrepancies This problem is generated by the discrepancy between words of different languages that express the same concept. These words can be referred to a singular pronoun in the source language and to a plural pronoun in the target language. In order to take into account number discrepancies in the generation of the pronoun into English or Spanish a set of morphological (number) rules is constructed. The left-hand side of the number rule contains the interlingua representation of the pronoun and the right-hand side contains the pronoun in the target language.

Gender Discrepancies Gender discrepancies came from the existing morphological differences between different languages. For instance, English has less morphological information than Spanish. The English plural personal pronoun *they* can be translated into the Spanish pronouns *ellos* (masculine) or *ellas* (feminine), the singular personal pronoun *it* can be translated into *él/éste* (masculine) or *ella/ésta* (feminine), etc. In order to take into account gender discrepancies in the generation of the pronoun into English or Spanish a set of morphological (gender) rules is constructed.

5 Evaluation of Generation Module

The generation module takes the interlingua representation as input. Previously, pleonastic *it* pronouns have been detected (with a **P**recision[2] of 88.75%), Spanish zero pronouns have been detected (89.20% **P**) and resolved (81.38% **P**), and anaphoric third person personal pronouns have been resolved in English and Spanish (80.25% **P** and 82.19% **P** respectively).

Once the interlingua representation has been obtained, the method proposed for pronominal anaphora generation into the target language is based on the treatment of number and gender discrepancies.

5.1 Pronominal Anaphora Generation into Spanish

In this experiment the generation of English third person personal pronouns into the Spanish ones has been evaluated.

We have tested the method on both literary and manual texts. In the first instance, we used a portion of the SemCor collection (presented in [7]) that contains a set of 11 documents (23,788 words) where all content words are annotated with the most appropriate WordNet sense. SemCor corpus contains literary texts about different topics (laws, sports, religion, nature, etc.) and by different authors. In the second instance, the method was tested on a portion of MTI[3] corpus that contains 7 documents (101,843 words). MTI corpus contains Computer Science manuals about different topics (commercial programs, word processing applications, device instructions, etc.).

We randomly selected a subset of the SemCor corpus (three documents –6,473 words–) and another subset of the MTI corpus (two documents –24,264 words–) as training corpus. The training corpus was used for improving the number and gender rules. The remaining fragments of the corpus were reserved for test data.

We conducted a blind test over the entire test corpus applying the number and gender rules. The obtained results appear in table 1.

Table 1 shows the anaphoric pronouns of each document classified by semantic roles: AGENT, THEME and MODIFIER. The last three columns represent the number of pronouns successfully resolved, the total number of pronouns resolved and the obtained **P**recision, respectively. For instance, the a13 document of the SemCor corpus contains 17 pronouns with semantic role of AGENT, 2 pronouns with semantic role of THEME and 3 pronouns with semantic role of MODIFIER. The **P**recision obtained in this document was of 95.45% (21/22).

[2] By **P**recision we mean the number of pronouns successfully resolved divided by the total number of pronouns resolved in the text. A detailed study of the evaluation of the different tasks carried out in order to obtain the interlingua representation in AGIR can be found in [12].

[3] This corpus has been provided by the Computational Linguistics Research Group of the School of Humanities, Languages and Social Studies –University of Wolverhampton, England–. The corpus is anaphorically annotated indicating the anaphors and their correct antecedents.

Discussion. In the generation of English third person personal pronouns into the Spanish ones an overall **P**recision of 80.39% (582/724) has been obtained. Specifically, 90.16% **P** and 75.11% **P** were obtained in SemCor and MTI corpus respectively.

Corpus		Subject	Complement		Correct	Total	P (%)
		AGENT	THEME	MODIF.			
SEMCOR	a02	21	5	1	23	27	85,19
	a11	10	5	0	14	15	93,33
	a13	17	2	3	21	22	95,45
	a14	40	10	1	48	51	94,12
	a15	32	5	4	34	41	82,93
	d02	14	2	3	18	19	94,74
	d03	13	0	1	12	14	85,71
	d04	50	6	9	59	65	90,77
	SEMCOR TOTAL	**197**	**35**	**22**	**229**	**254**	**90,16**
MTI	CDROM	38	24	7	47	69	68,12
	PSW	24	36	2	52	62	83,87
	WINDOWS	16	19	2	30	37	81,08
	SCANWORX	95	87	11	142	193	73,58
	GIMP	66	33	10	82	109	75,23
	MTI TOTAL	**239**	**199**	**32**	**353**	**470**	**75,11**
	TOTAL	**436**	**234**	**54**	**582**	**724**	**80,39**

Table 1. Generation of pronominal anaphora into Spanish. Evaluation phase

From these results we have extracted the following conclusions:

- In SemCor corpus all the instances of the English pronouns *he*, *she*, *him* and *her* have been correctly generated into Spanish. It is justified by two reasons:
  - The semantic roles of these pronouns have been correctly identified in all the cases.
  - These pronouns contain the necessary grammatical information (gender an number) that allows the correct generation into Spanish, independently of the antecedent proposed as solution by the AGIR system.

  The errors in the generation of pronouns *it*, *they* and *them* have been originated by different causes:
  - Mistakes in the anaphora resolution stage, i.e., the antecedent proposed by the system is not the correct one (44.44% of the global mistakes).This causes an incorrect generation into Spanish mainly due to the proposed antecedent and the correct one have different grammatical gender.

  - Mistakes in the identification of the semantic role of the pronouns that cause the application of an incorrect morphological rule (44.44%). These mistakes are mainly originated by an incorrect process of clause splitting.
  - Mistakes originated by the electronic dictionary from English to Spanish (11.12%). Two circumstances can occur: (a) the word does not appear in the dictionary; and (b) the word's gender in the dictionary is different to the real word's gender due to the word has different meanings.
- In MTI corpus, nearly all the pronouns are instances of the pronouns *it*, *they* and *them* (96.25% of the total pronouns). The errors in the generation of these pronouns are originated by the same causes than in SemCor corpus but with different percentages:
  - Mistakes in the anaphora resolution stage (22.86% of the mistakes).
  - Mistakes in the identification of the pronouns' semantic role (62.86%).
  - Mistakes originated by the English-Spanish dictionary (14.28%).

5.2 Pronominal Anaphora Generation into English

In this experiment the generation of Spanish third person personal pronouns (including zero pronouns) into the English ones has been evaluated.

We have tested the method on literary texts. We used a portion of the Lexesp corpus that contains a set of 31 documents (38,999 words). Lexesp corpus contains literary texts about different topics (politics, sports, etc.) from different genres and by different authors.

We randomly selected a subset of the Lexesp corpus (three documents –6,457 words–) as training corpus. The remaining fragments of the corpus were reserved for test data.

We conducted a blind test over the entire test corpus applying the number and gender rules. The obtained results appear in table 2.

Discussion. In the generation of Spanish third person personal pronouns into the English ones an overall **P**recision of 84.77% (657/775) has been obtained. From these results we have extracted the following conclusions:

- All the instances of the Spanish plural pronouns (*ellos*, *ellas*, *les*, *los*, *las* and zero pronouns in plural) have been correctly generated into English. It is justified by two reasons:
  - The semantic roles of these pronouns have been correctly identified in all the cases.
  - The equivalent English pronouns (*they* and *them*) lack gender information, i.e., are valid for masculine and feminine, then the antecedent's gender does not influence the generation of these pronouns.
- The errors occurred in the generation of the Spanish singular pronouns (*él*, *ella*, *le*, *lo*, *la* and zero pronouns in singular). They have been originated by different causes:
  - Mistakes in the anaphora resolution stage (79.66%).
  - Mistakes in the application of the heuristic used to identify the antecedent's semantic type (20.34%). This fact involves the application of an incorrect morphological rule.

Conclusion

In this paper the pronominal anaphora generation module of a complete interlingua MT approach (for Spanish and English languages) is presented and evaluated. The interlingua representation of the whole text is one of the main advantages of our system due to several problems, that are hardly solved by the majority of MT systems, can be treated and solved. These problems are the generation of intersentential anaphora, the detection of coreference chains and the generation of Spanish zero-pronouns into English. In the evaluation, the following results have been obtained: a Precision of 80.39% and 84.77% in the generation of English and Spanish personal pronouns (including zero pronouns) into the target language respectively.

Corpus		Subject	Complement		Correct	Total	P (%)
		AGENT	THEME	MODIF.			
LEXESP	txt1	19	3	1	21	23	91,30
	txt2	35	7	1	33	43	76,74
	txt3	21	4	1	19	26	73,08
	txt4	13	4	1	15	18	83,33
	txt5	13	4	1	14	18	77,78
	txt6	17	1	0	16	18	88,89
	txt7	22	3	4	28	29	96,55
	txt8	10	0	0	9	10	90
	txt9	9	3	1	8	13	61,54
	txt10	17	2	1	19	20	95
	txt11	7	0	1	7	8	87,5
	txt12	25	4	0	29	29	100
	txt13	16	0	0	12	16	75
	txt14	11	0	0	10	11	90,91
	txt15	16	3	5	18	24	75
	txt16	11	1	2	13	14	92,86
	txt17	14	1	0	11	15	73,33
	txt18	9	4	0	10	13	76,92
	txt19	7	0	1	7	8	87,5
	txt20	17	3	1	13	21	61,90
	txt21	4	2	0	6	6	100
	txt22	12	1	2	15	15	100
	txt23	15	4	2	19	21	90,48
	txt24	21	7	2	25	30	83,33
	txt25	92	11	5	100	108	92,59
	txt26	132	16	11	129	159	81,13
	txt27	24	6	1	27	31	87,10
	txt28	21	5	2	24	28	85,71
	TOTAL	**630**	**99**	**46**	**657**	**775**	**84,77**

Table 2. Generation of pronominal anaphora into English. Evaluation phase

References

1. L. Appelo and J. Landsbergen. The machine translation project Rosetta. In T.C. Gerhardt, editor, *I. International Conference on the State of the Art in Machine Translation in America, Asia and Europe: Proceedings of IAI-MT86, IAI/EUROTRA-D*, pages 34–51, Saarbrucken (Germany), 1986.
2. A. Berger, P. Brown, S.D. Pietra, V.D. Pietra, J. Gillett, J. Lafferty, R.L. Mercer, H. Printz, and L. Ures. The Candide system for Machine Translation. In *Proceedings of the ARPA Workshop on Speech and Natural Language*, pages 157–163, Morgan Kaufman Publishers, 1994.
3. J. Chandioux. MÉTÉO: un système opérationnel pour la traduction automatique des bulletins météreologiques destinés au grand public. *META*, 21:127–133, 1976.
4. R. Dale. *Generating referring expresions: constructing descriptions in a domain of objects and processes.* MIT Press, Cambridge, Mass, 1992.
5. A. Ferrández, M. Palomar, and L. Moreno. An empirical approach to Spanish anaphora resolution. *Machine Translation*, 14(3/4):191–216, 1999.
6. A. Ferrández and J. Peral. A computational approach to zero-pronouns in Spanish. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'2000)*, pages 166–172, Hong Kong (China), 2000.
7. S. Landes, C. Leacock, and R. Tengi. Building semantic concordances. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*, pages 199–216. MIT Press, Cambridge, Mass, 1998.
8. S. Lappin and H.J. Leass. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561, 1994.
9. G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. WordNet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244, 1990.
10. T. Mitamura, E. Nyberg, and J. Carbonell. An efficient interlingua translation system for multi-lingual document production. In *Proceedings of Machine Translation Summit III*, Washington, DC (USA), 1991.
11. C.D. Paice and G.D. Husk. Towards the automatic recognition of anaphoric features in English text: the impersonal pronoun "it". *Computer Speech and Language*, 2:109–132, 1987.
12. J. Peral. *Resolución y generación de la anáfora pronominal en español e inglés en un sistema interlingua de Traducción Automática.* PhD thesis, University of Alicante, 2001.
13. J. Peral and A. Ferrández. An application of the Interlingua System ISS for Spanish-English pronominal anaphora generation. In *Proceedings of the Third AMTA/SIG-IL Workshop on Applied Interlinguas: Practical Applications of Interlingual Approaches to NLP (ANLP/NAACL'2000)*, pages 42–51, Seattle, Washington (USA), 2000.
14. J. Peral and A. Ferrández. Generation of Spanish zero-pronouns into English. In D.N. Christodoulakis, editor, *Natural Language Processing - NLP'2000*, volume 1835 of *Lecture Notes in Artificial Intelligence*, pages 252–260, Patras (Greece), 2000. Springer-Verlag.
15. J. Peral, M. Palomar, and A. Ferrández. Coreference-oriented Interlingual Slot Structure and Machine Translation. In *Proceedings of the ACL Workshop Coreference and its Applications*, pages 69–76, College Park, Maryland (USA), 1999.
16. P. Vossen. EuroWordNet: Building a Multilingual Database with WordNets for European Languages. *The ELRA Newsletter*, 3(1):7–12, 1998.
17. A.P.M. Witkam. *Distributed language translation: feasibility study of multilingual facility for videotex information networks.* BSO, Utrecht, 1983.