영한기계번역과 대용어 조응문제에 대한 고찰

Ruslan Mitkov, 이강혁, 김형근, 최기선
전산학과
한국과학기술원

English-to-Korean Machine Translation and the Problem of Anaphora Resolution

Ruslan Mitkov, Kang-Hyuk Lee, Hiongun Kim, Key-Sun Choi
Computer Science Department
Korea Advanced Institute of Science and Technology (KAIST)

Abstract

At least two projects for English-to-Korean translation have been already in action for the last few years, but so far no attention has been paid to the problem of resolving pronominal reference and a default pronoun translation has been considered instead. In this paper we argue that pronouns cannot be handled trivially in an English-to-Korean translation and one cannot bypass the task of resolving anaphoric reference if aiming at good and natural translation. In addition, we propose lexical transfer rules for English-to-Korean anaphor translation and outline an anaphora resolution model for an English-to-Korean MT system in operation.

1. Anaphora Resolution and Mahine Translation

At least two English-to-Korean Machine Translation (MT) systems have been reported so far in operation ([Kim & Choi 93], [Lee & Kim 93]), but none of them has paid attention to the problem of resolving pronominal reference and default translation has been used for handling pronouns[1].

Everyone agrees that anaphora resolution is a complicated problem in natural language processing. Considerable research has been done by computational linguists ([Carbonel & Brown 88], [Dahl & Ball 90], [Frederking & Gehrke 87], [Hayes 81], [Hobbs 78], [Ingria & Stallard 89], [Rich & LuperFoy 88], [Robert 89]), but no complete theory has emerged which offers a resolution procedure with success guaranteed. Most approaches developed - even if we restrict our attention to pronominal anaphora, from purely syntactic ones to highly semantic and pragmatic ones, only provide a partial treatment of the problem.

Anaphora resolution within the domain of Machine Translation (MT) has its specific aspects. With an exception of a recent paper [Preuβ et al. 94] which offers an anaphora resolution model for English to German translation, there has been no work which investigates the anaphora resolution problems from the point of view of MT.

In MT, besides ambiguity resolution, often seen as its most important problem [Hutchins & Somers 92], another major difficulty is the resolution of anaphora[2].

The identification of pronouns involves the identification of the earlier noun phrases to which they refer, called the pronoun's antecedent[3]. The establishment of the antecedents of anaphora is very often of crucial importance for the correct translation. When translating into languages which mark the gender of pronouns for example, it is essential to resolve the anaphoric relation. Furthermore, the translation of the predicates connected with the pronoun (verbs, nouns etc.) may change according to different antecedents.

Consider a MT system with English as a source language and consider translating the pronoun "it" from English into the target language. If the target language is French, Spanish or Italian, the pronominal anaphoric reference has to be resolved before we decide which of the two possible pronouns - masculine or feminine - to use. In German, Greek and Slavic languages we have one more gender choice - neutral.

[1] In this paper we only consider pronouns in 3rd person singular and plural. Pronouns in 1st and 2nd person in Korean are socially dependant, rare in technical sublanguages and are not subject to our current investigations

[2] Given the complexity of the problem, we have concentrated on pronominal type of anaphora and later in our paper each reference of "anaphora" will be used as synonym of "pronominal anaphora".

[3] Also termed "referent" which is used as a synonym throughout the paper

In some languages the pronoun is translated directly by its referent. In English to Malay translation for instance, there is a tendency of replacing 'it' with its referent. Replacing a pronominal anaphor with its referent means, however, that the translator (program) must be able to identify first the referent.

Anaphora resolution reflects two essential topics in Machine Translation: ambiguity in a MT context and translation of discourse instead of isolated sentences. Anaphora can be viewed as a sort of ambiguity, in that the antecedent of a given pronoun might be uncertain and referential relations are one of the means that constitute coherence of texts.

2. Rationale for Anaphora Resolution in English-to-Korean Machine Translation

In Korean MT community, not much attention has been drawn to anaphora resolution problems. This is partly due to the complicated problem of anaphora resolution. But it is also due to the biased assumption that anaphoric expressions in the source language can be easily mapped to the corresponding anaphors in the target (Korean) language, or in many cases they can be simply ignored in the transfer phase.

Whereas in most European language pairs anaphora resolution is "compulsory" (or else we risk of rendering in certain cases quite unacceptable translations), there are certain cases in Korean where anaphora resolution may seem "optional".

Consider the sentences [Hutchins & Somers 92]:

(1) The monkey ate the banana because it was hungry.
(2) The monkey ate the banana because it was ripe.
(3) The monkey ate the banana because it was tea-time.

In each case the pronoun "it" refers to something different: in (1) the monkey, in (2) the banana and in (3) - to the abstract notion of time. If we have to translate the above sentences in German, then anaphora resolution is inevitable, since the pronouns take the gender of their antecedents and since the German words "Affe" - (masculine, "monkey"), "Banana" (feminine, "banana") and "es" (neutral - "it" for time notion) are in different gender.

Consider the translation of the sentences (1)-(3) from English-to-Korean and their literal descriptions in English.

(1') 배고파서 원숭이는 바나나를 먹었다.
hungry-CAUSAL monkey-NOM banana-ACC eat-PAST,DECL
(2') 익어서 원숭이는 바나나를 먹었다.
ripe-CAUSAL monkey-NOM banana-ACC eat-PAST,DECL
(3') 티타임이어서 원숭이는 바나나를 먹었다.
tea time-CAUSAL monkey-NOM banana-ACC eat-PAST,DECL

Note that in the above Korean translations there are no pronouns. These examples might seem encouraging that we could translate from English-to-Korean, bypassing the tough problem of anaphora resolution. However, such a conclusion would be too misleading.

The assumption that anaphoric expressions in the source language can be easily mapped to the corresponding anaphors in the target language, or in many cases they can be simply ignored in the transfer phase, is totally unfounded. It is not hard to find English sentences for which anaphora resolution is necessary in order to get their correct translation into Korean. Consider the sentences:

(4a) Although programmers usually write good programs, they may still make a mistake.
(4b) Although programs are usually written by good programmers, they may still contain mistakes.

In Korean, there are two types of pronominals corresponding to "they", one for human beings and the other for non-human beings. In order to assign the proper Korean pronominals to the English pronominal "they", the system should be able to resolve "they" between the two possible referents, "programmers" and "programs".

Anaphora resolution becomes a more serious business when we aim at achieving high-quality translation. The translation of (4a) and (4b) into Korean with the successful assignment of pronouns may still sound awkward to Koreans, because in Korean it is stylistically more natural not to explicitly mention anaphors in subordinate clauses that are coreferential with nominal expressions in the main clause. It is somewhat similar to English participle constructions whose subject is "understood." The best translation of (4b) in Korean could be described in English literally as:

(5) Being usually written by good programmers, programs may still contain mistakes.

Thus, if we are able to get the translation of (4a) and (4b) without overt pronominals, we are more likely to get better translation. This being so, anaphora resolution is very crucial in English-to-Korean MT because we must resolve the pronominal "they" to replace it by proper nominal expressions.

Moreover, "optional" anaphora resolution means preserving anaphoric ambiguity in case no anaphora resolution is undertaken. It may seem that carrying ambiguities over translation is even more "authentic" from the point of view of having a mirror translation of the source text. Not resolving anaphoric ambiguity means that during the translation process text is not fully understood. Generally speaking, however, analysis is aimed at producing an unambiguous intermediate representation [Isabelle & Bourbeau 85].

Moreover, a system strongly relying on the "ambiguity preservation" method, in addition to offering no computational advantage when ambiguity-preserving situations must be identified dynamically, is extremely vulnerable in situations where (i) the lexicon is growing while the system is in use or (ii) when additional languages must be introduced ([Nirenburg et al. 92]). Every new word sense added to the lexicon carries the potential of ruining the possibility of retaining ambiguity in translation for all previous entries. All this means that extra attention must be paid to the maintenance of the lexicons.

3. English-to-Korean Machine Translation and Mates

We are currently investigating anaphora resolution feasibility with regards to possible extension of our English-to-Korean translation system MATES/EK MATES/EK has been developed through a co-research done by KAIST and SERI (Systems Engineering Research Institute) from 1988 to 1992, and is still under evolution in KAIST. It is a transfer-based system, which does an English sentence analysis, transforms the result (parse tree) into an intermediate representation, and then transforms it into a Korean syntactic structure to construct a Korean sentence [Choi 94 et al].

- Morphological Analysis Using N-gram:

  Category ambiguities are resolved by combining the N-gram and the rules.

- Augmented Context Free Grammars for English Syntactic Analysis:

  An augmented context free grammar has been defined for general English syntactic analysis and the analyzer is implemented using Tomita LR parsing algorithm.

- Lexical Semantic Structure (LSS) to represent the intermediate representation:

  The result of the syntactic structure is transformed into an intermediate representation LSS, which is a dependency structure that is relatively independent to specific languages. In LSS, the constituents in a sentence are combined only in head-dependent relation based on the lexical categories, and there are no order relation between the constituents. Hence LSS is desirable for translation between English and Korean, two languages with fairly different syntactic structures.

- Grammar Writing Language and its Environment:

  MATES/EK runs a series of tree transformations on the LSS structure from the English syntactic structure, in order to get a structure specific to Korean syntactic structure. To do this, we developed a grammar writing language, in which the rules describe the tree transformations, and its supporting system. During the tree transformation operations the system looks up in the English-Korean bilingual dictionary in order to get the Korean lexemes.

- Development Supporting Tools :

  Taking into account the continuously growing property of an MT system, we developed a set of development supporting tools for grammar writing and editing, dictionary updating and translation testing.

MATES/EK consists of a set of dictionaries, a set of grammar rules and processing modules. Translation is carried out as sequential processing stages: English morphological analysis, English syntactic analysis, English semantic analysis, English-Korean lexical transfer, English-to-Korean structural transformation, Korean syntactic structure generation and Korean morphological generation. Figure 1 depicts the overall configuration of MATES/EK. MATES/EK has following features:

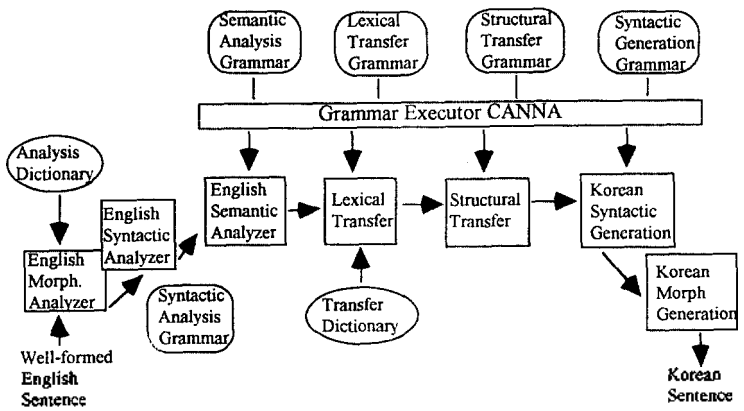

Figure 1. The system configuration of MATES/EK

This system is tested on 1,708 sentences, whose length is less than 26 words selected from 2500 sentences in the IEEE computer magazine September 1991. It shows about 95 percent of success rate for sentences shorter than 15 words, about 90 percent for 18 words, 80 percent for 21 words, and 75 percent for 26 words. This is a quite encouraging result since the magazine contains various kinds of texts of various styles.

4. Mates and Anaphora Resolution

Though MATES/EK has demonstrated very encouraging results, it is far from being a perfect automatic translator. We have not implemented an anaphora resolution module yet and there are various cases of complex sentences and/or discourse segments in which the system will not produce the most natural translation.

Consider the following two sentences and discourse segment:

(6a) The man who manages Tanya knows she is a difficult case.
(6b) Is animation really as useful as it seems?
(6c) Similarly to human beings, computers can share a multimedia vocabulary and common knowledge. They, however, can communicate with emotion.

These sentences could be translated correctly by MATES [Kim & Choi 93].

Now consider another two complex sentences (7a), (7b) and another discourse segment (7c) which have exactly the same syntactical structure as (6a), (6b) and (6c) respectively.

(7a) The president who governs Russia knows she is a difficult case.
(7b) Is the child really as hungry as it seems?
(7c) Similarly to human beings, computers can share a multimedia vocabulary and common knowledge. They, however, can be connected in an on-line network.

Due to the unavailability of an anaphora resolution model, wrong translations will be produced. The resolution of the anaphor "she" (non-human) in (7a) would require the translation "그것 (kugot)" (instead of the default "그녀 (kunyo)"), the resolution of "it" (human) in (7b) would help in assigning the correct translation "그애 (kuae)" (instead of the default "그것 (kugot)") and the resolution of the anaphor "they" (non-human) in (7c) would imply the use of "그것들 (kugotul)" (instead of the default "그들 (kutul)"). Moreover, a better translation is even possible (see rules 5.1-5.7).

It is clear that MATES needs an anaphora resolution module. But before coming finally to this problem, we had to investigate in detail the English-to-Korean anaphor translation phenomena.

5. Some Practical Lexucal Transfer Rules for English-to-Korean Anaphor Translation

We studied various texts from a Computer Science English corpus prepared for Machine Translation and on the basis of the observations we proposed practical anaphor translation rules for the needs of the lexical transfer in English-to-Korean Machine Translation. These rules prescribe how English anaphors should be translated into Korean on the basis of syntactic information, the type and semantic class of the noun the anaphor refers to, and are very useful for the lexical choice during the transfer. We do not regard the described set of rules as complete: however, it provides a good starting point and gives an idea of the direction of our current research efforts. The following major cases are concerned:

5.1 Complex Sentence I (main clause + subordinate clause)

When the subordinate clause in a complex sentence follows the main clause, the anaphor in the subordinate clause should not be translated.

(8a) Cashes are one of the most important ideas in computer architecture because they can substantially improve performance by the use of memory.
메모리 사용에 의해 성능을 상당히 향상 시킬 수 있으므로 캐쉬는 컴퓨터 구조에서 가장 중요한 아이디어중 하나다.
memory use-INSTRUMENTAL performance-ACCUSATIVE substantially improve-CAUSAL cache-NOMINATIVE architecture-LOCATIVE most important ideas-among one-DECLARATIVE.

(8b) The key is to partition computations so they rely on local data.
중요한 것은 로칼데이타에 의존하도록 계산을 나누는 것이다.
important thing-NOMINATIVE local data-DEPENDENT depend computation-ACCUSATIVE divide thing-DECLARATIVE.

5.2 Complex Sentence II (subordinate clause + main clause)

If the subordinate clause in a complex sentence is followed by the main clause, the anaphor(s) in the main clause is realized as the noun phrase(s) in the subordinate the anaphor refers to.

(9) As processors get faster, they will lose more and more of their performance to the memory system.
프로세서가 빨라질수록 프로세서의 메모리 시스템에 대한 성능이 상실될 것이다.
processor-NOMINATIVE fast-CONDITIONAL processor-POSSESSIVE memory system-DIRECTIVE performance-NOMINATIVE lose-FUTURE,DECLARATIVE.

5.3 Generalized Quantifiers

If the anaphor refers to generalized quantifiers[4], it is translated into definite description patterns such as "그런 (kuren: 'such' or 'that') + nominal".

(10) Although many approaches may be technologically feasible, they must also be economically feasible to be applied in a market economy.
많은 접근방식들이 기술적으로 적당하더라도 그런 접근방식들은 시장경제에 적용될 때 경제적으로 적당해야 한다.
many approaches-NOMINATIVE technically feasible-CONCESSIVE those approaches-NOMINATIVE market economy-LOCATIVE apply-MODIFIER time-TEMPORAL economically feasible-DECLARATIVE.

5.4. Human "it" (child)

If the antecedent of "it" is human, replace the pronoun with its antecedent (preceded by a definite article)

(11) The child is in the room, and it is playing with a doll.
아이는 방에 있는데, 그 아이는 인형을 가지고 놀고 있다.
child-NOMINATIVE room-LOCATIVE exist-CONJUNCTIVE, the child-NOMINATIVE doll-INSTRUMENT play-PRESENT,DECLARATIVE.

5.5 Non-human "she"

If the antecedent of "she" is non-human, translate the pronoun by replacing it with its its antecedent (preceded by a definite article) .

(12) There was a nice banquet on a ship. Her name was 'Pearl de Mer'.
배위에서 훌륭한 향연이 있었다. 그 배의 이름은 'Pearl de Mer' 였다.
ship-LOCATIVE nice banquet-NOMINATIVE be-PAST,DECLARATIVE. the ship-POSSESSIVE name-NOMINATIVE 'Pearl de Mer' be-PAST,DECLARATIVE.

5.6 Inanimate "they"

If the antecedent of "they" represents a set of non-human entities, translate the pronoun as "그것들 (kugotul)"

(13) When the system executes the erroneous instructions with certain data values, they cause a failure and the error becomes effective.
시스템이 어떤 데이타 값으로 오류가 있는 명령을 수행할 때, 그것들은 고장을 유발하고 오류가 효력을 발생한다.
system-NOMINATIVE certain data value-INCLUSIVE errors-NOMINATIVE exist-MODIFIER instructions-ACCUSATIVE execute-MODIFIER time-TEMPORAL, they-NOMINATIVE fault-ACCUSATIVE cause-CONJUNCTIVE error-NOMINATIVE effect-ACCUSATIVE occur-DECLARATIVE.

5.7 Every other case

In every other case, use the default translations : "he"-"그 (ku)", "she"-"그녀 (kunyo)", "it"-"그것 (kugot)" and "they"-"그것들 (kugotul)"

(14) Mr. Han went out at 3:00 in the afternoon and he will come back late in the evening.
한씨는 오후 3시에 나갔는데 그는 저녁 늦게 돌아올 것이다.
Mr.Han-NOMINATIVE afternoon 3:00-TEMPORAL go_out-PAST,CONJUNCTIVE he-NOMINATIVE evening late return-FUTURE,DECLARATIVE.

The rules 5.1-5.7 are to be integrated into the lexical transfer module of MATES. As one can see, these practical rules are based on available information about the referent and before applying them, antecedents should have been already identified.

The above rules are given in priority order and can be easily described in a more formal way.

6. Anaphora Resolution Model for Mates

In order to implement the above rules and since we are aiming at extending MATES into a system that can handle discourse translation (initially handling anaphoric references), we are studying different anaphora resolution strategies. At this stage, we have chosen a simplified version of our integrated anaphora resolution model proposed in [Mitkov 93]. Full implementation of this model, including center tracking inference engine, seems too costly for the immediate goals of our English-to-Korean translation system.

The main idea is that given the complexity of the problem, we think that to secure a comparatively successful handling of anaphora resolution one should adhere to the following principles: 1) restriction to a domain (sublanguage) rather than focus on a particular natural language as a whole; 2) maximal use of linguistic information integrating it into a uniform architecture by means of existing partial theories. Some more recent treatments of anaphora ([Carbonell & Brown 88], [Rich & LuperFoy 88]), ([Preuß et al 94]) do express the idea of "multi-level approach", or "distributed architecture", but their ideas a) do not seem to capture enough discourse and heuristical knowledge and/or b) do not concentrate on and investigate a concrete domain, and thus risk being too general. We have tried nevertheless to incorporate some of their ideas into our own proposals.

Our anaphora resolution model integrates modules containing different types of knowledge - syntactic, semantic, domain, discourse and heuristical (Figure 2).

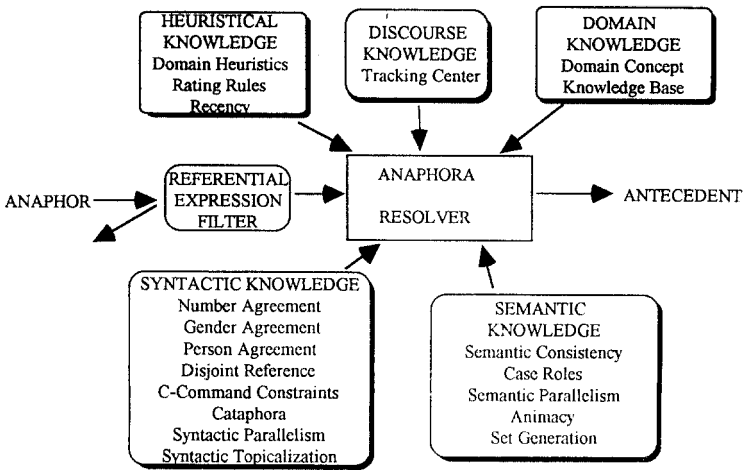

Figure 2: An integrated anaphora resolution architecture

The syntactic module, for example, knows that the anaphor and antecedent must agree in number, gender and person. It checks if the c-command constraints hold and establishes disjoint reference. In cases of syntactic parallelism, it prefers the noun phrase with the same syntactic role as the anaphor, as the most probable antecedent. It knows when cataphora is possible and can indicate syntactically topicalized noun phrases, which are more likely to be antecedents than non-topicalized ones.

The semantic module checks for semantic consistency between the anaphor and the possible antecedent. It filters out semantically incompatible candidates following the current verb semantics or the animacy of the candidate. In cases of semantic parallelism, it prefers the noun phrase, having the same semantic role as the anaphor, as a most likely antecedent. Finally, it generates a set of possible antecedents whenever necessary.

The domain knowledge module is practically a knowledge base of the concepts of the domain considered and the discourse knowledge module knows how to track the center of the current discourse segment.

The heuristical knowledge module can sometimes be helpful in assigning the antecedent. It has a set of useful rules (e.g. the antecedent is to be located preferably in the current sentence or in the previous one) and can forestall certain impractical search procedures.

The use of common sense and world knowledge is in general commendable, but it requires a huge knowledge base and set of inference rules. At the present stage of our project, however, we do not envisage the development of this module.

The syntactic and semantic modules usually filter the possible candidates and do not propose an antecedent (with the exception of syntactic and semantic parallelism). Usually the proposal for an antecedent comes from the domain, heuristical, and discourse modules. The latter plays an important role in tracking the center and proposes it in many cases as the most probable candidate for an antecedent.

The referential expression filter plays an important role in filtering out impersonal 'it'-expression (e.g. "it is important", "it is necessary", "it should be pointed out" etc.) , where 'it' is not anaphoric.

Initially, we envisage the implementation of the syntactic, semantic and heuristical modules, which together with the referential expression filter alone are helpful in solving practically most of the cases in our sublanguage.

References

[Barwise & Cooper 81] - Baswise J., Cooper R. - *Generalized quantifiers and natural languages*. Linguistics and Philosophy, 4, 1981

[Carbonell & Brown 88] J. Carbonell, R. Brown - *Anaphora resolution: a multi-strategy approach*. Proceedings of the 12. International Conference on Computational Linguistics COLING'88, Budapest, August, 1988

[Choi et al 94] Choi K.S. , Lee S.M. , Kim H.G. , Kim D.B, Kweon C.L., Kim G.C. - *An English-to-Korean Translator : MATES/EK*. Proceedings of the 15. International Conference on Computational Linguistics COLING'94, Kyoto, August, 1994

[Dahl & Ball 90] D. Dahl, C. Ball - *Reference resolution in PUNDIT*. Research Report CAIT-SLS-9004, March 1990. Center for Advanced Information Technology, Paoli, PA 9301

[Frederking & Gehrke 87] R. Frederking, M. Gehrke - *Resolving anaphoric references in a DRT-based dialogue system: Part 2: Focus and Taxonomic inference*. Siemens AG, WISBER, Bericht Nr.17, 1987

[Hayes 81] P.J. Hayes - *Anaphora for limited domain systems*. Proceedings of the 7th IJCAI, Vancouver, Canada, 1981

[Hobbs 78] J. Hobbs - *Resolving pronoun references*. Lingua, Vol. 44, 1978

[Hutchins & Somers 92] - *An Introduction to Machine Translation*, Academic Press, 1992

[Ingria & Stallard 89] R. Ingria, D. Stallard - *A computational mechanism for pronominal reference*. Proceedings of the 27th Annual Meeting of the ACL, Vancouver, British Columbia, 26-29 June 1989

[Isabelle & Bourbeau 85] Isabelle P. and L. Bourbeau - *TAUM-AVIATION: Its technical features and some experimental results*, Computational Linguistics, 11, 1985

[Kim & Choi 93] Kim C.G., Choi K.S. - *Machine translation environment system for English-to-Korean*, CAIR-TM-93-19, 1993

[Lee & Kim 93] - Lee H.G, Kim Y.T. - *An idiom-based approach to Machine Translation*. Proceedings of TMI'93, Kyoto, Japan, 14-16 July, 1993

[Mitkov 93] Mitkov R. - *A knowledge-based and sublanguage-oriented approach for anaphora resolution. Proceedings* of the Pacific Asia Conference on Formal and Computational Linguistics, Taipei, 30-31 August 1993

[Nirenburg et al. 92] Nirenburg S., Carbonell J., Tomita M., Goodman K. - *Machine Translation: a knowledge-based approach*, Morgan Kaufmann Publishers, 1992

[Preuß 94 et al] Preuß S., Schmitz B., Hauenschild C., Umbach C. - *Anaphora Resolution in Machine Translation*. In W. Ramm, P. Schmidt, J. Sch*1z (eds.) Studies in Machine Translation and Natural Language Processing, Volume on "Discourse in Machine Translation" (to appear)

[Rich & LuperFoy 88] E. Rich, S. LuperFoy - *An architecture for anaphora resolution*. Proceedings of the Second Conference on Applied Natural Language Processing, Austin, Texas, 9-12 February 1988

[Robert 89] M. Robert - *Resolution de formes pronominales dans l'interface d'interogation d'une base de donnees*. These de doctorat. Faculte des sciences de Luminy, 1989