Asian Journal of Information Technology 5 (1): 48-53, 2006

A Multiple Knowledge Sources Algorithm for Anaphora Resolution

Allaoua Refoufi
Département d'Informatique, Faculté des Sciences de l'Ingénieur Université de Sétif, Algeria

Abstract: This study presents a new algorithm for the resolution of anaphoric expressions which uses a few ideas largely accepted by the scientific community in computational linguistics as part of the resolution process. The algorithm driven by a robust morphosyntactic parser uses a set of constraints, which the candidates must not violate, as well as a set of preferences which discriminate among them. We also incorporate the notion of focused entities, and eliminate from further consideration entities included in insertions. The algorithm, written in Prolog, uses Definite Clause Grammars (DCG) for the implementation of the parser as well as the constraints and the preferences. Finally we discuss the results and the efficiency of the algorithm in light of related work in the field.

Key words: Anaphora, syntax, semantics, reference, focused entities

INTRODUCTION

The term anaphoraTPPT relates to the presence in the text of entities (noun phrases, pronouns, etc.) which, on one hand, refer to the same entity (are co referential) and, on the other hand supply additional information about these entities. Reference to an entity is generally termed anaphora, the entity to which the anaphora refers is the antecedent or the referent, anaphor is the entity used to make the reference. Example My brother called last night, he wants to see me he is the anaphor; my brother is the antecedent. In general several antecedents are possible for the same anaphor. This ambiguity constitutes the main challenge of the problem.

This study reviews the recent techniques for anaphora resolution and presents a new approach based on various syntactic, semantic, and pragmatic considerations discussed in the literature. A correct interpretation for anaphora is vital to the natural language processing systems: in information extraction, it allows to garnish the information structure with a complete detailed response to the initial reference. In text summarization, it allows to bring coherence to various parts of the input text. In translation it permits to choose the correct sense for the terms, eliminating therefore ambiguities. In opinion polls it allows to connect different topics to a unique entity. The process of finding the proper antecedent for each anaphora in texts is called anaphora resolution. Such resolution is very important because without it the text would not be fully and correctly understood.

In most cases, various expressions (noun phrases, pronouns, etc.) are either potential anaphor expressions, either potential antecedents. Anaphora resolution requires syntactic knowledge as well as semantic information (case roles, super classes, etc.) and makes use of pragmatic knowledge to convey information about focused entities which are essential for text comprehension. In any resolution method two phases are necessary: first localize anaphoric expressions, which designate entities elsewhere in the text, then find out the correct antecedent to each anaphora. Even the simplest anaphora expressions that the human encounters (and that we do not perceive consciously) require for their automatic treatment a considerable amount of knowledge. Despite considerable progress accomplished recently in this area of research, the solutions presented are relatively inefficient. We can achieve better performance by using different sources of knowledge at different levels, that is what we intend to show in this work.

The algorithm identifies noun phrase antecedents of personal, demonstrative, reflexive, and pleonastic pronouns in French. It identifies both intrasentential (when the anaphor and the antecedent occur in the same sentence) and intersentential (when the anaphor and the antecedent do not occur in the same sentence) antecedents and is applied to the output of the syntactic analysis generated by a robust parser.

The strategy combines different forms of knowledge and distinguishes between constraints and preferences. Whereas constraints are used as conditions that must not be violated, preferences sort the remaining candidates in an order which is believed to be optimal.

Main types of anaphora: We observe different types of anaphora in any natural language:

- Pronominal: it's the most used one, the reference is made by a pronoun «SabrinaB_{1B} prit [la pomme]B_{2B} sur la table. ElleB_1 laB_2 mangea » (Sabrina$_1$ took [the apple]$_2$ on the table. She$_1$ ate it$_2$)
- Definite noun phrase : the antecedent is referred to by a definite noun phrase [Le président]B_{1B} a visité la ville de Sétif .[L'hote du palais du people] B_{1B} inaugura plusieurs réalisations . ([The president]$_1$ visited the town of Setif. [The host of the people's palace]$_1$ inaugurated several realisations.
- Verb phrase as antecedent : Sarah essaya de [convaincre Rachid de se reposer]B_{1B} . [La tentative]B_{1B} fût vaine .(Sarah tried [to convince Rachid to rest]$_1$. [The attempt]$_1$ was vain)
- Ordinal Anaphora: ordinal anaphora is observed when the anaphor is a cardinal number like first, second, etc or an adjective such as former or latter. An example for ordinal anaphora could be: Sarah n'est pas satisfaite par [la solution]B_1 . Elle en chercha [une nouvelle]B_1 (Sarah was not satisfied by [the solution]$_1$. She looked for [a new one]$_1$.)
- One-anaphora is the case where the anaphor is one-phrase. J'ai acheté deux livres, le moins cher est le plus intéressant . (I bought two books, the cheapest is the more interesting.)
- Pleonastic pronoun (also known as semantically empty pronoun) is the case when a pronoun (usually il) does not refer to any particular antecedent. Phrases that use more words than semantically necessary are considered pleonastic. Examples are il pleut, il est important de noter que ... etc. pleonastic pronouns are not considered anaphoric (since they don't have an antecedent), identifying such occurrences is important so that the anaphora resolution system will not try to look for their antecedents.

The anaphor refers only to previous part of the text, contrary to cataphora which refers to entities not yet introduced. However, pronouns refer to entities evoked no further than two sentences back (whereas definite noun phrases, for example, can refer further back).

When performing anaphora resolution, all noun phrases are typically treated as potential candidates for antecedents. The scope is usually limited to the current and preceding sentences and all candidate antecedents within that scope are considered.

Most of the implementations for anaphora resolution employ a recency factor which states that if there are several candidate antecedents for an anaphor and all of these candidates satisfy the consistency restrictions for the anaphor (i.e. they are qualified candidates) then the most recent one (the one closest to the anaphor) is chosen.

Appositives, also termed insertions in the French literature, are usually used to provide some additional information for a named entity. The additional information is separated from the name of the entity by a comma and is usually placed immediately after the entity name. For example Ceasar, the roman emperor ... Appositional phrases are considered to supply additional information which is not of great interest in the text. The identification of appositives enables us to eliminate candidates which occur inside the apposition.

Sources of knowledge: Anaphora resolution requires multiple sources of knowledge. Morphology is concerned with the structure of words; it is almost inconceivable for a natural language application not to employ morphological knowledge. At the very least, applications require a lexicon or a dictionary. Morphological analysis tells us how to extract the base forms out of inflected forms that occur in texts. This type of analysis is especially important for French where inflected forms of verbs proliferate.

Syntax is concerned with the ways words combine to form phrases, and phrases combine to form sentences. Syntax associates some structure to the utterances of the language; moreover it tells us the syntactic function of each word (verb, noun, pronoun, etc.). This information is crucial to any anaphora resolution algorithm. The process of performing syntactic analysis is known as parsing. The process of parsing should be robust which means that it always terminates even in the presence of errors.

Another type of knowledge, semantics, let us know which phrases do make sense and which combinations should be rejected. It deals with the meaning of words, phrases and sentences. Even when a sentence is semantically unambiguous, it sometimes carries secondary meanings, depending on the context in which it is uttered. Pragmatic knowledge uses the context in order to disambiguate among different settings.

These types of knowledge are in fact hard to implement especially semantics and pragmatics; in fact the current state of art in these areas offers little insight about the true nature of interaction.

PREVIOUS WORK

Much study has been performed in the field of anaphora resolution and especially in the field of pronominal resolution. Works with significant importance include Lappin *et al*,[1-5]. Each method relies on various knowledge sources, we distinguish those that uses full syntactic parsers[1,5] from those that use only poor knowledge sources[2,3], which means only an output of a part of speech tagger that identifies only noun phrases and pronouns to be resolved. The fist three algorithms we review are for English texts, and the fourth one for French. Most often the algorithm for pronominal anaphora resolution consists of the following steps:

- Locate the anaphor(s) in the current input sentence.
- Reject the candidates that fail to satisfy consistency checks
- According to a predefined set of rules assign salience values to each candidate
- Choose the candidate ranked highest in the previous step

A pioneer study reported by Hobbs[6] uses a syntactic tree to search the input sentence for antecedents. The algorithm is a left to right breadth first search on the syntactic parse tree of the input sentence. Given the usual order of syntactic categories in the English language (the subject is generally followed by the verb) the algorithm expresses a preference for the noun phrases subjects.

Probably one the well known algorithm for pronoun resolution was proposed by Lappin and Leass[1]. The algorithm exploits salience factors and their associated weights such as sentence recency, subject emphasis, head noun emphasis), and so on to perform pronominal resolution. The salience value is simply the sum of the associated weights. Once salience values have been calculated for each referent, the algorithm can be applied to resolve the pronouns. The entity with the highest salience value is declared to be the most likely referent. If there are no pronouns to be resolved in a sentence, the next sentence is processed and the weights that contribute to an entity's salience are halved (to account for sentence recency). The weights used in the salience algorithm are ad hoc. Lappin and Leass's algorithm for pronominal anaphora resolution is capable of high accuracy, but requires in depth, full, syntactic parsing of text. The authors report 86% successfully identified antecedents in a corpus containing technical manuals.

Kennedy and Boguraev[2] describe a variant that does not require in-depth, full syntactic parsing of text. Instead, with minimal compromise in output quality, the modifications enable the resolution process to work from the output of a part of a speech tagger, enriched only with annotations of grammatical function of lexical items in the input text stream. Their method has been applied to personal pronouns, reflexives and possessives. The general idea is to construct co reference equivalence classes that have an associated value based on a set of ten factors. An attempt is then made to resolve every pronoun to one of the previous introduced discourse referents by taking into account the salience value of the class to which each possible antecedent belongs. The authors report 75.5% success in resolution on a corpus containing texts of different genres.

Based upon their own evaluation of the results of their implementation they state that accurate anaphora resolution can be realized within natural language processing frameworks which do not, or cannot, employ robust and reliable parsing components.

Mitkov's algorithm[3] is another knowledge poor approach to pronominal resolution, which means that it uses only the output of a part of speech tagger with minimal syntactic information. The algorithm does not employ syntactic information but relies on a set of indicators (rules) such as definiteness, heading, collocation, referential distance, term preference, etc. The indicators, boosting and impeding ones, assign salience values to the antecedents. The boosting indicators assign a positive score to a noun phrase, reflecting a positive likelihood that it is the antecedent of the current pronoun. In contrast, the impeding ones apply a negative score to a noun phrase, reflecting a lack of confidence that it is the antecedent of the current pronoun. A score is calculated based on these indicators and the discourse referent with the highest aggregate value is selected as antecedent. The author reports success rate of 89.7% on a corpus of technical manuals.

Trouilleux's algorithm[5] is a rule based pronoun resolution for French that uses full syntactic parsing. The method restrains the list of potential antecedents by using the notion of an insertion. He defines an insertion to be either a sequence between parentheses, a sequence delimited by commas between a verb and its subject or object, an apposition to the right of a noun phrase, or an apposition to the left of a subject noun phrase. The algorithm is reported to have a good rate of success (74,8%). The corpus used is newspaper articles in the finance domain.

Other study is concerned with implementing a pronominal resolution system and adjusting these weights empirically using machine-learning techniques based on real corpus data. Recently some authors use genetic algorithms to adjust the salience weights. Various corpora differ in the number of words, the domain of the texts, the types and the distribution of pronouns, types and distribution of named entities, the complexity of resolving certain constructs. An algorithm performing very well on a corpus of technical manuals may fail for a corpus of news articles or dialogs containing quoted speech. It is therefore important that the implementation considers specifics of the target texts upon which it is intended to operate.

THE RESOLUTION METHOD

Our system is composed of two main tasks: the first one, the recognition phase, performs parsing, recognition of non anaphoric pronouns, identification of focusing expressions and data structures building. The data structures built during parsing contain the main features of lexical items.

The second part, the anaphora resolution procedure, applies the constraints and the preferences. A constraint defines a property that must be satisfied in order for any candidate to be considered as a possible solution of the anaphor. The constraints used in the algorithm are the following: morphological agreement (gender, number and person) and conditions on insertions (apposition). A preference is a characteristic that is not always satisfied by the solution of an anaphor. The aim of the preferences is to obtain a ranked list of candidates. Some examples of preferences used in our system are the following: syntactic parallelism, antecedent non included in a prepositional phrase, focused expressions and recency.

The parser written in Prolog is based on a set of definite clauses, which are grammar rules augmented with arguments used to capture features and build structures. Structures building are derived from the parse trees generated in the firs part of the algorithm

Non anaphoric expressions are signalled by expressions of the form il est { possible, évident, admis, normal, pertinent, logique, courant, etc.} que … and indicate that the pronoun il is not anaphoric. The number of template matching used to identify such constructions can be updated or augmented.

Focussing is defined as the process which chooses a theme, or center of attention, in a discourse and moves it as the speaker's discourse proceeds. The focus provides a valuable source for identifying pronominal anaphora. Focusing expressions are of the form c'est NP qui, il y a NP, where the noun phrase NP is the center of focus..

Resolution algorithm
Recognition phase:

- Morphosyntactic analysis
- Recognition of non anaphoric pronouns
- Identification of focusing expressions
- Data structures building

Resolution phase: For each anaphor do :
Until one antecedent is found do:

- Carry out in order the constraints
- Carry out in order the preferences

The constraints: Constraints are rules which participate in the purging of the candidates appearing in the structures built during the parsing process. Incorporated constraints are:

- Consistency conditions
- Condition on insertions: an expression which is included in an insertion cannot be the antecedent of an anaphor located outside the insertion.

Consistency conditions are agreement on morphological grounds (gender, number and person). The insertion constraint stipulates that in the example : [Le mariage]B_{2B} de SabrinaB_{1B} sœur cadette de Sofia et Sarah, aura lieu la semaine prochaine. ElleB_{1B} s'yB_{2B} prépare activement , the candidates Sofia and Sarah, are to be eliminated from further process.

Preferences: Preferences, as opposed to constraints, can be violated by the antecedent candidates, they are used to rank the candidates. However those that verify the preferences are retained. The order in which they appear reflects their weight. The preferences incorporated are :

- Syntactic parallelism
- Antecedent not occurring in a prepositional phrase
- Focused expressions
- Recency

Syntactic parallelism states that we prefer the antecedent that shares the same syntactic function as the anaphor. For example. : L'enfantB_{1B} reconnut le roiB_{2B}. pourtant ilB_{1B} ne lB_{2B}'avait jamais rencontré auparavant . The antecedent le roi is discarded.([The child]$_1$ recognized [the king]$_2$ although he$_1$ has never met him$_2$ before.)

An expression, mainly a noun phrase, included in a prepositional phrase is unlikely to be referred to because it only brings additional information. For example in the sentence : [La voiture]B_{1B} de la voisine nous bloque le chemin, il faut laB_{1B} déplacer the anaphor la refers to la voiture and certainly not to la voisine ..

The recency preference favours the candidate which lies nearest to the anaphor, that is the one evoked recently.

DISCUSSION

The syntactic parser used has been under construction for the last two years at the University of Sétif, it is however far from being completely achieved. Major improvements will be made on the semantic level, we aim to incorporate essential semantic information that can be used for major natural language applications.

For the time being the parser has a wide linguistic coverage of French syntax, and it uses a dictionary of about 600 words. Proper names are only accepted if they belong to a special list, which can be updated easily. In particular named entities or unknown words still provoke failures.

The external mistakes that have negative impact on the anaphora resolution module performance are: not recognized named entities and sentences not properly parsed.

Our objective is to process the following subject pronouns (il, ils, elle, elles), the object pronouns (l', le, la, les), possessive determinants (son, sa, ses, leur, leurs). For possessive determinants we look for the antecedent inside the current sentence, only when the procedure fails we make an attempt outside the current sentence. For example SamirB_{1B} jette sonB_{1B} chapeau sur la table

Our system does not recognize multiple source anaphor: when the anaphor refers to several antecedents at the same time. An illustration of this phenomena is given in the sentence Sarah et Sofia sont parties tôt ce matin. Elles ont rendez vous à l'université .(Sarah and Sofia left early this morning. They have an appointment at the university)

Self referring expressions are not dealt with. An example of this type of anaphora might be Chacun le sait, Rachid est un as du volant . (Everyone knows it, Rachid is a good driver)

When using the pronoun l' we face more difficulties because the gender feature is missing. We also encountered the case where the gender feature of the pronoun elle is of little help in the determination of the correct antecedent, as in the example Le docteur Bouzidi est un spécialiste en chirurgie. Elle travaille jour et nuit. (Doctor Bouzidi is a specialist in surgery. She works day and night).

The pronoun lui can be either a masculine or feminine indirect object, as in the example elle lui donne la clé .

Our algorithm does not deal with anaphors that refer to verb phrases or sentences, as in : Sur un deux roues, on est très fragile. Le problème c'est de l'oublier . (On two wheels we are fragile. The problem is to forget it).

The algorithm realises a success rateTPPT of 68%, the corpora used is extracted from literary textbooks where the phenomena of anaphora is very dense, as opposed to scientific texts. Texts used consist of 3 to 5 chapters, each chapter contains about 5 sentences, each sentence contains 5 to 20 words. The evaluation has been carried out so far on 36 texts of reasonable size.

The results show that the resolution of pronouns such as il(s), elle(s) is relatively successful (success rate of 93%). The algorithm will also benefit from some syntax information indicating the subject of the sentence, because the results show that the recency factor and the gender agreement are not sufficient.

Our system is capable of identifying two types of appositions: those located between parentheses or brackets, and those located between commas. The implementation for the latter apposition is a bit

complex, as we have to make sure that the final comma does exist before deciding whether we are in the presence of an apposition.

The insertion constraint tends to add more complexity in the implementation, which to the best of our knowledge does not carry real improvement to the algorithm.

Our method does not use salience measures to discriminate among competing candidates, because we argue that salience measures are randomly fixed and hence do not bear any scientific value

CONCLUSIONS

The main idea of our work consists of the establishment of a link between nominal phrases that share similar context with constituents in the input text. The method relies heavily on a morphosyntactic robust parser, where the main knowledge is gathered. The application of a set of constraints followed by a set of preferences provides an elegant modular, easy to update anaphora resolution algorithm. We believe that only a full syntactic parsing can provide the required knowledge to the anaphora resolution module. Unfortunately, current state-of-the-art of practically applicable parsing technology still falls short of robust and reliable delivery of syntactic analysis of real texts to the level of detail and precision that most algorithms assume. Shallow parsing, on the other hand, can affect greatly the performance and the efficiency of the algorithm. In comparison with previous work, the success rate realised (68%) is very satisfactory, and motivates us to explore new ways in which to improve our work.

REFERENCES

1. Lappin S. and H.J. Leass, 1994. An algorithm for pronominal anaphoric resolution, computational linguistics, 20: 535-561.
2. Kennedy, C. and B, Boguarev, 1996.Anaphora for everyone: Pronominal anaphora resolution without a parser. Proceedings of the 16th International Conference on Computational Linguistics (COLING'96),Copenhagen, Denmark pp: 113-118.
3. Mitkov, R., 1998. Robust pronoun resolution with limited knowledge, Proceedings of the 18th International Conference on Computational Linguistics, Montreal, Canada, 1998
4. Mitkov, R., 2001. Outstanding issues in Anaphora Resolution, in A. Gelbkh(Ed), in Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing (CICLing), Mexico City.
5. Trouilleux, F., 2002. Insertions et interprétations des ex^pressions pronominales. In Actes de l'Atelier, Chaînes de référence et résolveurs d'anaphores TALN 2002, Nancy.
6. Jerry R.H., 1978. Resolving pronoun references. Lingua, 44: 339-352.