Lappin and Leass' Algorithm for Pronoun Resolution in Portuguese*

Thiago Thomes Coelho and Ariadne Maria Brito Rizzoni Carvalho

Institute of Computing, State University of Campinas,
Mail Box 6176, 13084-971 Campinas (SP), Brazil
{thiago.coelho, ariadne}@ic.unicamp.br

Abstract. This paper presents a variant of Lappin and Leass' Algorithm for pronoun resolution in Portuguese texts; the algorithm resolves third person pronominal anaphora, as well as reflexive and reciprocal pronouns. It relies on salience measures, derived from the syntactic structure of the sentence, and on a simple discourse representation model. The algorithm, as well as its evaluation with legal and literary corpora, are presented.

1 Introduction

The phenomenon of coreference that occurs in natural language is the device of making an abbreviated reference to some entity (or entities) in the expectation that the perceiver of the discourse will be able to disabbreviate the reference and, thereby, determine the identity of the entity. The abbreviated reference is called an anaphor and the entity to which it refers is its referent or antecedent. The process of determining the referent of an anaphor is called resolution [1].

Anaphora resolution can improve significantly the performance of several natural language processing applications, such as automatic translation and summarisation, among others. The main difficult is to identify the proper referent when there exists more than one candidate. Several approaches to anaphora resolution have been proposed, such as Lappin and Leass Algorithm [2], Hobbs Algorithm [3] and Centering Algorithm [4].

Lappin and Leass' Algorithm, or RAP (*Resolution of Anaphora Procedure*), aims at identifying both intra and intersentential antecedents of third person pronouns and lexical anaphors (reflexive and reciprocal), in English. This paper presents a variant of Lappin and Leass' Algorithm for pronoun resolution in Portuguese, as well as its evaluation with legal and literary corpora. This algorithm was chosen because of its good performance with English texts.

The remainder of this paper is organized as follows: in the next section the Lappin and Leass' original algorithm is presented; in section 3 the algorithm built for pronoun resolution in Portuguese is described and an example of its execution is presented; in section 4 the algorithm is evaluated on two different corpora and the results are shown; and finally, in section 5 the conclusions and future work are presented.

* The work was partially sponsored by CNPq.

C. Bento, A. Cardoso, and G. Dias (Eds.): EPIA 2005, LNAI 3808, pp. 680–692, 2005.

2 Lappin and Leass' Algorithm

The main components of the Lappin and Leass Algorithm are [2]:

- An intrasentential syntactic filter [5,6] for ruling out anaphoric dependence of a pronoun on a noun phrase, based on syntactic grounds;
- A morphological filter for ruling out anaphoric dependence of a pronoun on a noun phrase due to person, number, or gender non-agreement;
- An anaphor binding algorithm [6] for identifying the possible antecedent binder of a lexical anaphor (reciprocal or reflexive pronoun) within the same sentence;
- A procedure for assigning the suitable salience factors weights to a noun phrase, according to its grammatical role, such as syntactic parallelism, subject, etc.;
- A decision procedure for selecting the preferred element from a list of possible antecedent candidates.

The anaphor binding algorithm identifies the intrasentential candidates for reflexive or reciprocal pronouns; the syntactic filter rules out the intrasentential coreference candidates for third person pronouns that are unlikely to be the antecedent. For the remaining candidates, the value of the salience factors are calculated (as described in section 2.1). The chosen referent will be the one with the highest salience factor. When there is more than one candidate with the same salience factor, the algorithm chooses the candidate which is closer to the pronoun. The syntactic filter and the anaphor binding algorithm analyse the pronoun's sentence syntactic structure to decide if coreference is allowed. The algorithm uses the grammatical representation generated by the parser developed by McCord [7,8].

2.1 Salience Factors

RAP uses a salience weighting system based on syntactic features. There are two types of operations performed by the algorithm: discourse model update and pronoun resolution. When a noun phrase, which introduces a new entity in the discourse is found, a representation for that entity is created and its salience factor is calculated. The salience factor for a given entity is the total of all salience factors applied to that entity. The initial salience factors are presented in Table 1.

Table 1. Salience factors with initial weights [2]

Salience factors	Weights
Sentence recency	100.0
Subject emphasis	80.0
Existential emphasis	70.0
Accusative emphasis	50.0
Indirect object and oblique complement emphasis	40.0
Non-adverbial emphasis	50.0
Head noun emphasis	80.0