A Characterization of Underspecified Anaphora and its Consequences on the Annotation of Anaphoric Relations

Frédéric Landragin

Underspecification'07

November, 2nd, 2007

Content

1. Objectives of the study

2. Partial interpretation and anaphora: phenomena and characterization
   - semantic ambiguity *versus* NLP ambiguity
   - ambiguity for the hearer *versus* ambiguity assumed by the speaker
   - a first classification of potential vague antecedents
   - discussion

3. Anaphora with vague antecedent and automatic anaphora resolution
   - annotation for evaluation *versus* annotation for representation
   - some annotation principles dedicated to vague antecedents
   - double mark-up principle and MIN attribute
   - discussion

4. Conclusion and perspectives

Objectives

1. Showing that it is not always useful to determine the precise antecedent of an anaphora

2. Bringing the notion of “anaphora with fuzzy / vague antecedent” to analyze a kind of partial interpretation (underspecified, ambiguous)

3. Studying the impact of discourse entity semantics on the identification of the set of potential antecedents for an anaphora

4. Providing recommendations for anaphora resolution systems, with the proposition of some annotation principles

Vague antecedent: case 1

1. « Je **lui** ai dit sur un ton de plaisanterie que **son idée** était intéressante, qu'**elle** montrait les choses sous un angle […] » Frantext K899

   "I told her that her idea was interesting, that **she-it** shows things in an unexpected way"

   - possible antecedent 1 = **the idea**, that shows things in an unexpected way
   - possible antecedent 2 = **the person**, that shows things in an unexpected way
   - two entities, one of them being a facet / property / function of the other
   - in French, both are possible (gender OK, number OK…) → **semantic ambiguity**
   - does the ambiguity trouble the comprehension of the text?
   - is the ambiguity assumed by the speaker?

2. « […] dans **le comportement de tout public**, quel qu'**il** soit »

   "the behavior of the public, whatever **he-it** is"

   Other examples: « une commune acception de la réalité […] **elle** […] »
   « la nostalgie d'une innocence perdue […] **elle** »
   « le caractère de Marcello […] **il** »
   « Marcello […] son engagement […] **il** »

Vague antecedent: case 2

1. « **les rapports des rapporteurs et les documents auxquels ils se réfèrent**, pour autant qu'**ils** n'ont pas été communiqués » Frantext P659

   "the reports from the reviewers and the related documents, if **they** have not been […]"

   - possible antecedents = first coordinate, second coordinate, both
   - coordination, especially with plurals → **semantic ambiguity**
   - does the ambiguity trouble the comprehension of the text?
   - is the ambiguity assumed by the speaker?

2. « **cette faille, cette malédiction** qui est comme une résurgence laïque du péché originel, est d'autant plus grave qu'**elle** laisse un vide […] »

   "this flaw, this curse is all the more serious since **it** leaves a void […]"

   - correction or reformulation: the antecedent can be: 1. the last term, 2. both terms
   - precision: the antecedent can be: 1. the last term, 2. both terms, (3. the first term)
   - apposition: the antecedent is preferentially the first term
   - same phenomenon of **semantic ambiguity**
   - *NB: « le rêve, l'évasion est nécessaire. **Il** permet de mieux supporter… »*
     *« le rêve, l'évasion est nécessaire. **Elle** permet de mieux supporter… »*

Vague antecedent: case 3

1. « Voulez-vous **Jean Martin pour époux** ? – Oui, je **le** veux »;
   « Jean a **coupé tout le bois**. Il **l'**a fait en petits morceaux »

   "do you want to take John Martin to be your husband? – Yes, I want **him-it**";
   "John cut wood. He did/made **it** in small parts"

   - possible antecedents = "wood", "cut wood"
   - **semantic ambiguity between an individual anaphora and an event anaphora**
   - does the ambiguity trouble the comprehension of the text?
   - is the ambiguity assumed by the speaker?

2. « **Jean s'est cassé la jambe. Il n'est pas allé au cinéma. C'**est arrivé hier »

   "John has broken his leg. He did not go to the cinema. **It** happened yesterday"

   - possible antecedents = 1. the first event,
     2. the set of two events (with the causal relation)
   - **semantic ambiguity between two event anaphora**

Types of ambiguities

1. Semantic ambiguity and NLP ambiguity
   - previous cases: semantic ambiguities because potential antecedents correspond to different discourse entities, with different semantic properties (*focus on the semantic elements*)
   - delimitating the antecedent for an abstract anaphora: is the antecedent of "**it**" the complete sentence, or only its verbal head? → NLP ambiguity (*focus on textual elements*)

2. Ambiguity for the hearer and ambiguity assumed by the speaker
   - semantic ambiguities: ambiguity for the hearer, and eventually also for the speaker
     - ambiguity for the hearer but probably not for the speaker: "I told her that her idea was interesting, that **she-it** shows things […]"
     - ambiguity for both interlocutors but assumed by the speaker: "this flaw, this curse is all the more serious since **it** leaves a void […]"

Types of vague antecedents

1. Possessives, when:
   - the possessor is animated and the possession is a personality trait to which it can be assimilated (facet, property, function…)
   - the possessor is an object and the possession is its main function
2. Complex nominal phrases "the N_1 of the N_2" with similar participants
3. Coordinations
   - with similar participants
   - in cases where the sentence elements are in plural: "**the N_1** and **the N_2** … **they**"
4. Juxtapositions
   - with similar participants
   - when it is hard to distinguish a simple enumeration from a reformulation, and, in the latter case, to distinguish a precision from a correction ("this flaw, this curse…")
   - example: "the First Secretary, Mr. Smith, and his wife" when the text does not allow to determine whether Mr. Smith is the First Secretary or a third person

Vague antecedents: discussion

1. Comprehension and anaphora with vague antecedent
   - close to the point of view of (Prandi 1987) concerning some syntactic ambiguities: “this kind of structural ambiguity has **no consequences on the interpretation**, maybe because it is frequent, systematic in some situations, and cannot be canceled”
   - among all the occurrences of anaphora in corpus, anaphora with vague antecedent are not very frequent…

2. As a statement
   - this notion of underspecified anaphora is an argument pro underspecification and partial interpretations: we do not need to analyze all anaphoric relations to understand a text
   - one aim of this study is to model (and formalize) this kind of partial interpretations

Annotating: why?

1. Annotation as a methodological phase
   - annotation as a step in the conception cycle of a model:

   - annotation for the evaluation of anaphora resolution systems
   - annotation for comparing systems

2. Annotation as a convenient means to represent interpretations
   - annotate allows to force oneself to delimitate potential antecedents, and therefore to better apprehend semantic phenomena
   - various interpretation hypotheses can be presented in a simple and clear manner using annotation
   - annotation is not dependent from the semantic theory or formalism from the concerned system, which is an argument pro genericity and favor the exchanges in the scientific community

Some annotation principles

1. Grouping principle

   grouping some antecedents into a composite one:
   "**John** was sleeping. **Mary** was reading. **They** were happy" {John, Mary}

2. Alternatives principle

   the initial intended antecedent from the speaker is in the set of alternatives, all of them being relevant for the hearer (and for the semantic representation):
   "I told **her** that **her idea** was interesting, that **she-it**…" {her | her idea}

3. Feasibles principle

   set of possible antecedents with a syntactic point of view, some of them (and not all) being relevant for the semantic representation determination, and some of them being cause of mistake: "the First Secretary, Mr. Smith, and his wife"
   (2 ou 3 persons, one of these feasibles being contrary to the reality)

4. Double mark-up principle

   inferior and superior limits for the text span that includes the antecedent

Concerning the double mark-up

1. For abstract anaphora like "**it** happened yesterday"
   - the alternatives principle is not sufficient, because the alternatives may be too numerous: "**it**" can refer to the whole previous sentence, to a part of it, and sometimes to several previous sentences or to the whole previous paragraph
   - corpus examples with the same kind of NLP ambiguity:
     "in **that context**", "in **this respect**", "to do **that**"
   - the best solution would be to manage fuzzy or progressive limits for the antecedent, which is not compatible with (XML) mark-up constraints

2. MIN attribute and double mark-up principle
   - MIN attribute from MUC-7: allows to define the minimal text span (minimal mention like a proper noun) in a mark-up that delimitates a longer text span (maximal)
     → same discourse entity
   - INF mark-up and SUP mark-up: allow to define the limit inferior and the limit superior of the text span where the vague antecedent is
     → cover several discourse entities
   - a MIN attribute can be added to the INF mark-up as well as to the SUP mark-up

Discussion on annotation

1. Towards an annotation schema
   - from the annotation principles to a complete annotation schema
   - is it really useful? what about systems that prefer to ignore some ambiguities and therefore be more efficient than systems that identify all ambiguities?

2. Towards an anaphora resolution system for underspecified anaphora
   - since the categories of vague antecedents are linked to syntactic criteria, an algorithm for underspecified anaphora resolution seems to be feasible
   - but is it really useful? vague antecedents are not so frequent in corpus, then taking them into account will not significantly increase the abilities of a system

3. As a statement
   - one aim is to provide some ideas to automatic solvers designers
   - another aim is to encourage approaches that take into account the partial, ambiguous, underspecified character of linguistic interpretation

Conclusion and perspectives

1. Presentation of a first attempt to collect data on underspecified anaphora and to propose a characterization of this phenomenon
   - for now my data are limited, because it is very hard to question the Frantext corpus on nominal phrases like "the N_1 of the N_2" or on text spans like "the N_1, the N_2 … they": too many irrelevant results, too many calculations, etc.
   - for now my characterization is more syntactic than semantic…
2. Perspectives
   - go ahead with the collect of data, explore other methods to question corpus
   - go deeper into the study of the links between N_1 and N_2:
     what are the semantic properties that link them?
     what are the constraints and the consequences on the interpretation?
   - model and formalize underspecified anaphora within an underspecified semantics framework