Back to Part I

3. Anaphoric Annotation: Simple Issues

3.1 Markable identification

Most current schemes focus on nominal anaphora to antecedents introduced by nominals, and require coders to mark all and only NPs, and to mark their full span. However particularly in the language of news we encounter rather complex nominals, as in (3.1.1).

(3.1.1) [Simón Bolívar, the mercurial Venezuelan aristocrat who led [South America's] [[19th-century] wars [of independence]]

Apart from the question whether appositions should be grouped together with the head of the NP -- here *Simón Bolívar* -- or be labelled as a separate entity, there is often the need to label embedded constituents. Some of them are referential *(South America)*, others may be not *(independence).*

Challenge for annotation tools We want to be able to comfortably label such embedded constituents, ideally with a graphical tool, and ideally on the basis of a syntactic tree: text-based tools make such annotations cumbersome. SALTO is a possibility but it produces output in SalsaXML.

Challenge for the guidelines Which modifying expressions are referential?

- Depending on the (syntactic) domain used for the annotation, there may be different referents (and thus potentially different labels) expressed by the same phrase.
- E.g. a prepositional phrase refers to a different entity than the noun phrase which is a part of the PP: [*underneath* [*the table* →object] → location] If *the table* has been mentioned before, the location can still be said to be (r-)new, whereas the object is (r-)given.

A separate issue is that of evaluation of markable identification. Marking the full span considerably penalizes systems unless MIN-IDs are also annotated, but not all schemes require annotators to specify MIN-IDs. Also many tools do not directly support that – e.g., MMAX. The solution adopted in SemEval 2010 was to automatically extract

MIN-IDs from the gold dependency trees, but this requires gold dependency trees (constituency trees are not as good).

Question Is there any reason not to mark MIN-IDs other than cost / limitations of the tools?

Challenge for tools at least with MMAX, annotating MIN-IDs is tricky

Challenge for guidelines how to specify the MIN-ID, or head, for coordination and for discontinuous cases.

3.2 Discourse new/discourse old

We know that discourse-new mentions can reliably be distinguished from discourse-old mentions (Poesio & Vieira, 1998), at least in the 'clearest' cases, e.g., anaphoric reference to people or locations using proper names or pronouns that do not refer to stages of individuals (but see 'quasi-coreference' below).

How we actually decide that two (individual type) entities are identical in reference may still be an unsolved puzzle from both a philosophical and a processing perspective (see Haug's comments, p.1) but in most cases we -- including the annotators -- seem to be fairly good at it.

3.3 Relations

Most current schemes distinguish between predication and identity (see, e.g., !AnCora/ARRAU/GNOME/OntoNotes), but the guidelines for marking predication tend to be purely syntactic: cases in which the predication can be inferred like (3.3.1) are typically not marked.

(3.3.1) Don't be fooled by Fluffy's name, lions are dangerous.

Some schemes also require the annotation of the few cases of bridging reference that has been shown to be relatively easy to mark (e.g., Set/element, some cases of part-of, other anaphora), but the general case of bridging reference is harder (see Section 5).

3.4 Singletons

In corpora like OntoNotes only multi-mention entities are annotated, but there are a number of advantages in annotating singletons as well—e.g., it can be helpful for training anaphoricity detectors, and provides a better test for systems running on gold mentions. So some schemes do require coders to annotate singletons (e.g., GNOME, ARRAU).

Challenge for the Guidelines Is there any reason not to mark singletons (other than saving annotation time)?

3.5 Referring and Non-Referring NPs

Some schemes (e.g., ARRAU, GNOME, ?Potsdam commentary corpus?) require annotators to mark non-referring NPs, but not all do (e.g., AnCora, OntoNotes).

Challenge for the Guidelines Is there any reason not to mark non-referring NPs?

In ACE, many NPs that in linguistics are generally not considered referring are treated as such and marked as coreferent, like the two mentions of *nobody* in (3.5.1).

(3.5.1) Nobody is scared of Fluffy. Nobody understands his secret powers.

Other schemes that do ask coders to mark non-referring NPs also require them to specify the semantic type of the NP (e.g., referring, expletive, predicative) -- e.g., ARRAU. Such schemes however tend not to do very well with quantifiers.

Challenge for the Guidelines What should we do with quantifiers?

Finally, most schemes we are aware of do not say much about a number of more complex types of NPs including for instance measure NPs as in (3.5.2), or NPs in comparative constructions as in (3.5.2).

(3.5.2) You need *two spoonfuls of sugar* to make the medicine go down.

(3.5.3) *The older* John got, *the sillier* he became.

3.6 Zeros

Pretty much all languages allow for phonetically unrealized arguments (zeros / traces) at least in non-tensed clauses, and many such as Catalan, Czech, Italian, Japanese, Spanish, etc. allow them in many more positions.

(3.6.1) Giovanni arrivo’ tardi. Ø Si scuso’ immediatamente con l’ospite.

There isn’t much theoretical disagreement concerning such elements (although some syntactic theories do question their existence), but they do raise quite a lot of issues with annotation and resolution. Coding such arguments is only possible with annotation tools that use as base layer a full syntactic annotation or on argument structure, such as TrEd used for coreference annotation in Prague Dependency Treebank; but even when such tools are available, corpora annotated in this way are problematic for many systems as the state of the art at extracting full syntactic structure / argument structure from text is still not good enough. And anyway many annotation tools only allow tokens as the base layer. With tools such as MMAX, the only option is to have ‘verbal markables’ as done e.g., in the VENEX/!LiveMemories annotation (i.e., annotate a relation between the immediately following verbal element and the antecedent). In AnCora, zero subjects were added as extra 'empty' tokens, and these were used for annotating coreference. A slightly cleaner solution is available with tools allowing character standoff.

Challenge for the Guidelines From the guidelines point of view, the question is which of unrealized arguments should be annotated (see below). But with tools that depend on token standoff one also needs to specify how to identify these zeros and how to mark verbal markables.

Challenge for Annotation Tools Is there still any technological constraint on developing annotation tools that rely on argument structure as their base layer?

Challenge for coreference resolvers Is the state of the art in dependency parsing improved enough to have a decent recall on zeros? At least one some of them? (E.g., recall for subject zeros in Japanese is OK, but on object zeros is pretty bad. (Iida & Poesio, 2011))

3.7 Grammatical Anaphors

Anaphoric annotation in corpora such as AnCora or the Prague Dependency Treebank includes grammatical anaphoric relations (ie the relations that are normally represented using coindexing in syntax papers) that would not be annotated in, e.g., ARRAU, GNOME, or OntoNotes. Examples of these relations include the relation between a relative pronoun (or relative trace) and the coindexed NP as in (3.7.1) and (3.7.2), or the relation between traces and other non-realized elements and their coindexing element (control relations, etc), as in (3.7.3). (This would also require addressing the zero problem, see Section 3.6).

(3.7.1) The guest who arrived late apologized

(3.7.2) John, who arrived late, apologized.

(3.7.3) John asked Mary Ø to come.

In ARRAU/GNOME/etc. only non-restrictive relatives are marked as in (3.7.2); the reason was to economize effort since that they ought be marked in the syntactic layer (although in practice they weren't).

Challenge for the Guidelines With modern corpora in which more and more there are multiple annotation layers, is there still any reason to annotate syntactic coindexing relations?

3.8 Premodifiers

NP modifiers can include both proper names, as in (3.8.1), and nouns, as in (3.8.2)

(3.8.1) "When the lights go up, it feels like the building is part of the street" SFJAZZ founder Randall Kline said in November.

(3.8.2) Open bars poured craft beer and wine cocktails

Some schemes require coders to mark mentions such as SFJAZZ in (3.8.1) as markables (e.g., ARRAU, GNOME, OntoNotes), but most (?) schemes do not require to treat wine in (3.8.2) as a markable, with the exception of ARRAU, which however (i) requires such modifiers to be marked as generic (see below) and (ii) only requires coders to treat these mentions as markables when they are subsequently referred to, which makes matters very difficult for systems attempting to do mention identification. On the other end, systematically marking all premodifiers as mentions requires substantial additional work.

In part the matter is related to the question of whether premodification acts as an anaphoric island, as proposed e.g., by Postal, and questioned by a number of people.

Challenge for the Guidelines Should these premodifiers always be treated as markables? Never? Only sometimes? Are there any clear linguistic tests that could be used to identify referring premodifiers?

3.9 Coordination

Coordination and its connection with plural reference are another difficult issue for work on empirical anaphora and information structure. The semantics of plural reference to a coordination like John and Mary met. They had not seen each other in a long time. is fairly uncontroversial from a semantic point of view, but an annotation project in principle could follow at least two different strategies:

1. Create a single markable for the entire coordination. Advantages: less annotation effort; entire coordination may serve as antecedent to a plural pronoun as in the

example above. Disadvantage: some other mechanism is needed for cases like (3.9.1).

2. Assign separate markables to each conjunct/disjunct, and then use a split antecedent mechanism to handle plurality. Advantages: the conjuncts may have different information status; each conjunct may corefer with other markables as in (3.9.2); antecedents consisting of disjoint markables (aggregation) seem necessary anyway as in (3.9.1). When augmented with a mechanism for discontinuous markables as in MMAX, can also handle cases of coordinated heads in which one of the coordinates is subsequently referred to as in (3.9.3). Disadvantages: no system we are aware of can deal with either split antecedents or discontinuous markables.

(3.9.1) John visited Ellen, and they went to the seaside.

(3.9.2) For his birthday, John got [a skateboard and a book]. He read the book in one night.

(3.9.3) The King and Queen of Hearts were sitting on their throne when Alice appeared. The Queen said severely “Who is she?”

Challenges for the guidelines what should we do with (3.9.3)? How about disjunctions or other types of coordination, as in

(3.9.4) Alice had no idea what [Latitude] was, or [Longitude] either, but thought they were nice grand words to say.

(Notice that, again, this case presents no semantical difficulty we are aware of –it’s the annotation that’s the problem.)

Challenge for annotation tools tool must provide a possibility to either split an antecedent link in order to cover discontinuous antecedents, or to link to subparts of complex antecedents.

Challenge for anaphoric resolvers how to deal with split antecedents / discontinuous markables?

3.10 Context Shifts

When you annotate narrative texts such as the New Testament, a major problem is that the text shifts between the narrative itself and direct speech. The narrative can refer to referents introduced in the direct speech, but the direct speech cannot refer to referents outside its own context. Here are two examples.

(3.10.1) Now after that John was put in prison, Jesus came into Galilee, preaching the gospel of the kingdom of God, And saying, "The time is fulfilled, and the kingdom of God is at hand: repent ye, and believe the gospel". (Mark 1:14-15)

(3.10.2) And she brought forth her firstborn son, and wrapped him in swaddling clothes, and laid him in a manger; because there was no room for them in the inn. And there were in the same country shepherds abiding in the field, keeping watch over their flock by night. And, lo, the angel of the Lord came upon them, and the glory of the Lord shone round about them: and they were sore afraid. And the angel said unto them, "Fear not: for, behold, I bring you good tidings of great joy, which shall be to all people. For unto you is born this day in the city of David a Saviour, which is Christ the Lord. And this shall be a sign unto you; Ye shall find the babe wrapped in swaddling clothes, lying in a manger". (Luke 2:7-12)

The second occurrences of "gospel" and "manger" corefer with the first ones, but since they are in direct speech, they cannot really be anaphors: how could Jesus' public and the shepherds in the field know the preceding narrative? That is why the angel prefers the indefinite article in Luke 2 -- in the dialogue situation, he is introducing a new referent. In the first example, we find a definite article, but this (so we assume) is due to Jesus relying on his audience being able to find the reference of gospel through general knowledge. The solution to this problem adopted in **project** has been to dissociate the IS tags from the anaphoric links. Such mentions are marked as coreferent with the previous mention, but also as ACC-gen/NEW. So we get NPs that are NEW but do have an anaphoric link.

Challenge for the guidelines Is this reasonable?

Go to Part III

RAIS/DocumentPartII (zuletzt geändert am 2013-04-18 15:14:37 durch ArndtRiester)

- IMS home
- wiki-support(at)ims
- Legal Notice