Anaphoric Annotation and Information Structure: Solved Problems, Open Challenges

Massimo Poesio & Arndt Riester

Collating text provided by: Stefan Baumann, Aoife Cahill, Christian Chiarcos, Philippa Cook, Kordula de Kuthy, Dag Haug, Yufang Hou, Katja Markert, Detmar Meurers, Anna Nedoluzhko, Malvina Nissim, Marta Recasens, Michael Strube, Olga Uryupina, Ramon Ziai

1. Introduction

This document is our attempt to collate the analyses of the state of the art in the annotation of anaphora and information structure produced by ourselves and the other participants to the workshop, identifying ‘solved’ and ‘open’ issues. It is based on the assumption that the two annotation tasks are related (one of us believes that anaphoric annotation is a subset of the annotation of information structure, the other remains agnostic).

2. State of the Art

2.1 Anaphoric Annotation

The first substantial anaphora annotation effort was the UCREL corpus developed by Uni Lancaster for IBM (Fligelstone, 1992). The guidelines for this annotation were actually quite ambitious covering also bridging relations and ellipsis, but we are not aware of any reliability study, and the corpus was never made available. The next widely known effort was the MUC annotation, whose guidelines are described in (Chinchor and Hirschman, 1997). The developers of the MUC scheme did carry out some form of reliability testing, deciding as a result to concentrate on what was called the ‘coreference’ relation which includes anaphoric identity and predication (Deemter & Kibble, 2000). The guidelines also contained detailed instructions regarding markable identification and

introduced the notion of ‘MIN-ID’. The next major proposal in the area, the DRAMA scheme (Passonneau, 1997) was proposed by the Discourse Resource Initiative. This scheme is much more ambitious and linguististically much more advanced than the MUC scheme (e.g., it specifies guidelines for marking bridging references) but it was never used for a major annotation effort. These early schemes and other ones were surveyed as part of the efforts of developing the MATE toolkit, a generic annotation tool meant to support annotation of coreference as well as of other discourse and semantic levels such as prosody and dialogue acts. The MATE ‘meta-scheme’ was a markup scheme that could have supported any of these types of annotation (Davies & Poesio, 1999).

The work on schemes supporting information extraction continued with the development of the ACE scheme (Doddington et al, 2002), which introduced an important innovation, entity-based annotation, and which was used for the annotation of the ACE corpora, until recently the main resources for studying coreference.

On the linguistically oriented annotation front, the MATE markup was adopted more or less directly for the GNOME annotation (Poesio, 2004). The GNOME guidelines were based on systematic reliability testing. All NPs were considered markables. The annotators were required to identify the semantic type of an NP (non referring, referring, predicative, quantifier, coordination) prior to annotate anaphoric information. Only anaphoric relations to antecedents introduced by NPs were marked (i.e., no abstract anaphora, and no event anaphora). A limited range of associative relations was also marked. The subsequent ARRAU guidelines were developed on the basis of a series of experiments testing the annotation of ambiguity and reference to abstract objects, among others. Ambiguity and reference to abstract objects were also central to the guidelines developed for the Potsdam commentary corpus (Krasavina & Chiarcos 2007). In recent years, linguistically-oriented schemes have been used for all major annotation efforts, including AnCoRa (Recasens & Marti), COREA (Hendrickx & Hoste), LiveMemories (Rodriguez et al, 2010), OntoNotes (Pradhan et al) and Tuba/DZ (Versley et al) (see (Poesio et al, in preparation) for discussion).

2.2 Information Structure

Information structure theory deals with pairs of opposite notions like focus-background, theme-rheme, topic-comment etc. Finding definitions of these pragmatic concepts which are applicable in linguistic annotation has turned out to be a challenge. Focus, for instance, is often defined as the answer to a 'question under discussion' (Roberts 1996), which, however, is typically not directly observable in monological text. Similarly, topic has sometimes been characterised as the cognitively most salient entity in a sentence, which necessitates an explanation how linguistic cues relate to the fact that an entity becomes salient in the minds of the interlocutors.

Approaches to the annotation of the main information structural notions have mainly proceeded in three ways: (i) rely on certain linguistic hypotheses that, e.g. words which carry a pitch accent (in spoken discourse) or which occur in certain syntactic positions necessarily mark a focus or a topic, (ii) devise question-answer tests to identify the focus constituent, or (iii) trace these pragmatic notions back to more primitive ones for which we have better intuitions.

As for focus, one such "primitive" system (not intended pejoratively) is the one by Schwarzschild (1999), which is based on a technical definition of the notion of givenness. On Schwarzschild's system, focus must occur on, or within, constituents which are not given. Givenness, in turn, is defined differently for referring and for non-referring constituents. For referring expressions, givenness is identical to coreference. In that sense, the annotation of coreference anaphora (Section 2.1) is an important building block of information structure annotation. For non-referring expressions (like verbs/verb phrases, adjectives, common nouns etc.), givenness means being entailed by an earlier constituent (i.e. by repetition, synonymy, hypernymy). (The two ways of being given are also found in Halliday and Hasan, 1976.) Baumann and Riester (2012) provide an annotation system (RefLex), which is based on these two notions of givenness.

A different annotation tradition which is closely intertwined with both anaphora and information structure is *information status* (Prince 1981, 1992). It springs from the insight that grouping referring expressions into 'anaphoric' and 'non-anaphoric', or into 'given' and 'new' ones is unsatisfactory. Not only is it oversimplistic; it can also cause confusion because certain expressions may have properties related to both givenness and newness, for instance, deictic expressions on their initial mention, bridging anaphora, or generic expressions occurring repeatedly. Several classification proposals have been made made in the literature which all seem to share some of the basic insights but which offer vastly different solutions with regard to terminology, definition of classes, and hierarchical organisation (Prince 1981, 1992; Gundel et al. 1993; Lambrecht 1994; Eckert & Strube 2000; Nissim et al. 2004; Götze et al. 2007; Riester et al. 2010; Baumann & Riester, 2012).

The notion of aboutness topic and the partition of utterances into a topic-comment articulation has become popular and aboutness topic considered to be a valid and important notion in much work on information structure within theoretical linguistics; as the overview in Krifka's (2008) "Basic Notions of Information Structure" makes clear. The basic notions that Krifka defines there have come to be considered standard by many researchers within the (Berlin-Potsdam) SFB 632 which focuses specifically on information structure. The problem of actually operationalizing the notion of aboutness topic is, however, not well-known and is discussed in Cook & Bildhauer (2011) and in an extension of that study reported on in Cook & Bildhauer (to appear). In the first

study, the authors undertook to annotate (from a set of potential referents) what was the aboutness topic of each sentence, using data extracted from a corpus of German newspaper texts. Although the annotators (= the authors) would consider themselves knowledgeable about information structure, the degree of inter-rater agreement was disappointing. In the later study, students were given training using the guidelines for annotation of information structure created within SFB 632, viz. Götze et al (2007). They then had to select the aboutness topic in each sentence (from a given set of markables), using texts from a different corpus of German newspaper articles. Again, the results were disappointing.

2.3 Computational models

Statistical models of anaphora resolution and of information structure annotation are at different stages of development, but neither is able to carry out even the limited range of interpretive tasks that can currently be annotated.

For instance, even though modern anaphoric annotation tools can be used to annotate cases in which the antecedent of a plural expression is not a single component (as in *John saw Mary. THEY greeted each other*), we are not aware of any anaphoric resolver able to produce an interpretation for such cases. Drawing on developments within anaphora resolution and named-entity recognition, as well as on the explosive grow of machine learning technology in NLP, there have been recent attempts to apply similar techniques in order to automatically classify referring expressions into two (anaphoric, new), three or many information status categories. Existing annotated corpora have been used as training data. Now, we would like to gain an overview on which features and which external resources have been employed, and to what effect. The same problems that hamper the comparison of manual annotations (how to map different classifications, unreliable annotation of certain categories) also cause trouble for the comparison of computational models.

Computational models of the classification of information status (coarse grained, fine grained) seem to be in a rather good shape. A few research groups reported quite some progress in 2011 and 2012 (e.g. Rahman & Ng; Cahill & Riester; Markert et al.). A few categories in the fine grained classification task still cause trouble, e.g. perhaps the most important "mediated/bridging" subcategory in Markert et al. (ACL 2012). The dependence on gold information (e.g. mentions, syntax, entity types) is bothersome.

Go to Part II

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