a subset of each document (e.g. interrogative sentences) without looking at other sentences.

More recently Martschat et al. (2015) introduced error analysis for mention pair types in the CORT system, which keeps track of each pair of mentions corresponding to a correct or incorrect linking decision in a mention-chain model.[9] For example, it is possible to diagnose precision or recall errors involving a pronominal anaphor with a common noun-headed antecedent, by counting correct and incorrect links of this type, in much the same way used by the MUC metric.

Building on Martschat et al.'s insights, we extend the MUC metric to features of single mentions involved in correct or incorrect links. We call this metric 'p-link', which stands for 'partitioned link score'. The basic idea is that a coreference failure (or success) has two equally responsible mentions in a consecutive mention-chain model. Each of the two mentions involved shares credit or blame for the classification decision. If a link partition is worth 1 precision or recall point, then involvement in a correct decision earns 0.5 points for the category that includes the mention at each end of the link.

Figure 1 illustrates this using the example from Pradhan et al. (2014), which has been extended with shading representing categories.

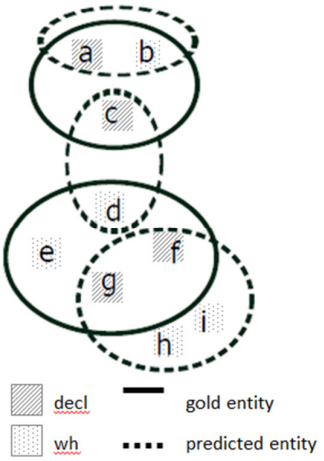

Figure 1: Gold (solid) and predicted (dashed) entities, with mentions in two categories distinguished by shading.

The solid oval represent two gold entities, with mentions {a,b,c} and {d,e,f,g}. Dashed ovals give three predicted entities, with mentions {a,b}, {c,d} and {g,f,h,i}. Note that mention e is not in any predicted entity, and h+i are not in the gold data. Pradhan et al.'s implementation of the MUC metric tallies the partitions with respect to gold and predicted mentions, such that a predicted link a+b is a correct positive (since a+b are in the same gold entity), c+d is a false positive, and the absence of predicted b+c is a false negative.

The p-link score builds on this by counting 0.5 points of correct positive, correct negative, etc. for each mention, such that points accrue for the respective category of that mention. The metric is a direct extension of Pradhan et al.'s definitions for recall (R) and precision (P):

$$p\text{-}link_{R,\pi} = \frac{\sum_{i=1}^{N_k}(|K_i^\pi| - p(K_i^\pi))}{\sum_{i=1}^{N_k}(|K_i^\pi| - 1)}$$

$$p\text{-}link_{P,\pi} = \frac{\sum_{i=1}^{N_r}(|R_i^\pi| - p'(R_i^\pi))}{\sum_{i=1}^{N_r}(|R_i^\pi| - 1)}$$

where K_i is the i^{th} entity in the key (gold) data (and Ri is correspondingly the i^{th} response entity); $|K_i^\pi|$ is the weighted partition magnitude within entity i, i.e. the number of instances of a mention from partition type π being either the source or target of a coreference link, multiplied by the weight 0.5 (since source and target may be of different types, and each is worth 'half a link'); and $p(K_i^\pi)$ is the set of elements of type π obtained by intersecting the key entities with the response entities, with each mention again being worth 0.5 points for its respective type π.[10]

Thus for the example in Figure 1, declaratives get 0.5 points for their correct involvement in a+b, but none for the missing link with c, and 1 point for their involvement in the correct g+f (since both are *decl*). The total possible links for declaratives in Figure 1 are worth 2 points (0.5 for a+b, 0.5 for b+c and 1 for g+f), so that *decl* scores a recall of 1.5/2 or 0.75 in this example. Indeed, only 1 of 4 *decl* link endpoints is missed in this example. We have implemented the p-link metric as an extension to Pradhan et al.'s original code, and our code is freely available.[11]

To test whether genre or sentence type has more influence on p-link, we evaluate manual and automatic coreferencer output, using a con-

[9] This approach assumes a 'mention-pair' model, in which each anaphor is linked to its antecedent in a chain. By contrast, 'mention-cluster' or 'entity-mention' models (see Rahman & Ng 2011) focus on entities as clustered groups of mentions referring to the same entity.

[10] Although we assign anaphors and antecedents equal weights of 0.5, other weights are conceivable.

[11] Code available at: `https://github.com/amir-zeldes/reference-coreference-scorers`.