Data

The goal of OntoNotes coreference annotation and modeling is to fill in the coreference portion of the shallow semantic understanding of the text that OntoNotes is targeting. For example, in “She had a good suggestion and it was unanimously accepted", we mark a case of IDENT coreference (identical reference) between “a good suggestion” and “it”, which then allows correct interpretation of the subject argument of the “accepted” predicate.

Names, nominal mentions, and pronouns can be marked as coreferent. Verbs that are coreferenced with a noun phrase can also be marked as IDENT; for example “grew” and “the strong growth” would be linked in the following case: “Sales of passenger cars grew 22%. The strong growth followed year-to-year increases.”

In order to keep the annotation feasible at high agreement levels, only intra-document anaphoric coreference is being marked. Furthermore, while annotation is not limited to any fixed list of target entity types, noun phrases that are generic, underspecified, or abstract are not annotated.

Attributive NPs are not annotated as coreference because the meaning in such cases can be more appropriately taken from other elements in the text. For example, in “New York is a large city”, the connection between New York and the attributive NP “a large city” comes from the meaning of the copula “is”. Similarly, in “Mary calls New York heaven”, the connection comes from the meaning of the verb “call”. Thus these cases are not marked as IDENT coreference.

Appositive constructions are marked with special labels. For example, in “Washington, the capital city, is on the East coast”, we annotate an appositive link between Washington (marked as HEAD) and “the capital city” (marked as ATTRIBUTE). The intended semantic connection can then be filled in by supplying the implicit copula.

Task

The coreference resolution task will use the English language portion of the OntoNotes data, which consists of a little over one million words from newswire (□ 450k), magazine articles (□ 150k), broadcast news (□ 200k), broadcast conversations (□ 200k) and web data (□ 200k). The task would be to automatically identify coreferring entities and events given predicted

information on the other layers. Preliminary investigations using this data were presented in Pradhan et al. (2007). Since OntoNotes coreference data spans multiple genre, we will be creating a test set spanning all the genres.

It is customary for CoNLL tasks to have two modes -- *open* and *closed*. The former allows for almost unrestricted use of external resources to complement the provided data so as to gauge the ceiling on the state of the art, while the latter is used to allow a fair comparison of the algorithms using the distributed data alone. We will similarly have an *open* and *closed* mode.

Closed

In the *closed* mode, the systems will be required to use only the provided data. We will provide the underlying text, and *predicted* versions of all the supplementary layers of annotation, during testing, using off-the-shelf tools such as parser, name entity tagger, etc. For the training data however, we will also provide gold-standard annotations of all the layers. Systems can either use the gold-standard or predicted annotation for training their systems. They can also train their own models for the various layers of annotation for which we provide predicted annotations to possibly get more accurate predicted information -- both during training and testing. However, they should restrict the tools they train, to the training portion of the data, and the annotation types in it. Unlike previous CoNLL tasks, the task of coreference requires world knowledge, and most state-of-the-art systems use information from resources such as WordNet, to add a layer of semantics that allows them to generalize connections between various lexicalized mentions in the data. Therefore, we propose to allow the use of WordNet (Word senses in OntoNotes are effectively coarse-grained groupings of WordNet senses) as part of the closed task. Another significant piece of information that is particularly useful for coreference and is not available in the layers of OntoNotes is that for number and gender. There are many different ways of generating this information, so in order to continue with the tradition that the systems in the *closed* task are restricted to the provided dataset for better algorithmic comparisons, we will provide an automatically pre-computed list of number and gender information values for the training and test text.

Open

In the *open* mode, in addition to the above, systems can use external resources such as Wikipedia. An advantage of the *open* mode is that participants might be able relatively easily to modify existing research systems in order to participate in the task thereby

easing the barrier to participation.

Evaluation

The OntoNotes data distinguishes between identity coreference and appositive coreference. We will evaluate the systems on the identity coreference task which links all categories of entities and events together into equivalent classes. Unlike MUC, or ACE, the OntoNotes data does not explicitly identify the minimum extents of an entity mention, so for the official evaluation, we will consider a mention to be correct only if it matches the exact same span in the annotation key . Since this data has hand-tagged syntactic parses, we plan to use the head words of the gold syntactic parses along with the extents to perform a more relaxed evaluation where a mention would be considered correct if the span falls within the span of the key mention and it contains the head word of that key mention. The evaluation of coreference has been a tricky issue and there does not exist a silver bullet, so we propose to compute a number of existing scoring metrics -- MUC, B-CUBED, CEAF and BLANC -- following the approach used by Recasens et al. (2010) . In spite of the variety of metrics, we do need to compute a single score to determine the winning system. After a lot of deliberation we have decided to use the unweighted average of MUC, B-CUBED and CEAF scores to determine the winning system. Only identity coreference (IDENT) will be considered in both tasks.

We will also evaluate systems on *anaphoric* mention detection since OntoNotes has tagged only mentions that have a coreferent mention in the same document. Singleton mentions were not annotated.