CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes

Sameer Pradhan
BBN Technologies,
Cambridge, MA 02138
pradhan@bbn.com

Lance Ramshaw
BBN Technologies,
Cambridge, MA 02138
lramshaw@bbn.com

Mitchell Marcus
University of Pennsylvania,
Philadelphia, 19104
mitch@linc.cis.upenn.edu

Martha Palmer
University of Colorado,
Boulder, CO 80309
martha.palmer@colorado.edu

Ralph Weischedel
BBN Technologies,
Cambridge, MA 02138
weischedel@bbn.com

Nianwen Xue
Brandeis University,
Waltham, MA 02453
xuen@cs.brandeis.edu

Abstract

The CoNLL-2011 shared task involved predicting coreference using OntoNotes data. Resources in this field have tended to be limited to noun phrase coreference, often on a restricted set of entities, such as ACE entities. OntoNotes provides a large-scale corpus of general anaphoric coreference not restricted to noun phrases or to a specified set of entity types. OntoNotes also provides additional layers of integrated annotation, capturing additional shallow semantic structure. This paper briefly describes the OntoNotes annotation (coreference and other layers) and then describes the parameters of the shared task including the format, pre-processing information, and evaluation criteria, and presents and discusses the results achieved by the participating systems. Having a standard test set and evaluation parameters, all based on a new resource that provides multiple integrated annotation layers (parses, semantic roles, word senses, named entities and coreference) that could support joint models, should help to energize ongoing research in the task of entity and event coreference.

1 Introduction

The importance of coreference resolution for the entity/event detection task, namely identifying all mentions of entities and events in text and clustering them into equivalence classes, has been well recognized in the natural language processing community. Automatic identification of coreferring entities and events in text has been an uphill battle for several decades, partly because it can require world knowledge which is not well-defined and partly owing to the lack of substantial annotated data. Early work on corpus-based coreference resolution dates back to the mid-90s by McCarthy and Lenhert (1995) where they experimented with using decision trees and hand-written rules. A systematic study was then conducted using decision trees by Soon et al. (2001). Significant improvements have been made in the field of language processing in general, and improved learning techniques have been developed to push the state of the art in coreference resolution forward (Morton, 2000; Harabagiu et al., 2001; McCallum and Wellner, 2004; Culotta et al., 2007; Denis and Baldridge, 2007; Rahman and Ng, 2009; Haghighi and Klein, 2010). Various different knowledge sources from shallow semantics to encyclopedic knowledge are being exploited (Ponzetto and Strube, 2005; Ponzetto and Strube, 2006; Versley, 2007; Ng, 2007). Researchers continued finding novel ways of exploiting ontologies such as WordNet. Given that WordNet is a static ontology and as such has limitation on coverage, more recently, there have been successful attempts to utilize information from much larger, collaboratively built resources such as Wikipedia (Ponzetto and Strube, 2006). In spite of all the progress, current techniques still rely primarily on surface level features such as string match, proximity, and edit distance; syntactic features such as apposition; and shallow semantic features such as number, gender, named entities, semantic class, Hobbs' distance, etc. A better idea of the progress in the field can be obtained by reading recent survey articles (Ng, 2010) and tutorials (Ponzetto and Poesio, 2009) dedicated to this subject.

Corpora to support supervised learning of this task date back to the Message Understanding Conferences (MUC). These corpora were tagged with coreferring entities identified by noun phrases in the text. The de facto standard datasets for current coreference studies are the MUC (Hirschman and Chin-

Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task, pages 1–27,
Portland, Oregon, 23-24 June 2011.

chor, 1997; Chinchor, 2001; Chinchor and Sundheim, 2003) and the ACE[1] (G. Doddington et al., 2000) corpora. The MUC corpora cover all noun phrases in text, but represent small training and test sets. The ACE corpora, on the other hand, have much more annotation, but are restricted to a small subset of entities. They are also less consistent, in terms of inter-annotator agreement (ITA) (Hirschman et al., 1998). This lessens the reliability of statistical evidence in the form of lexical coverage and semantic relatedness that could be derived from the data and used by a classifier to generate better predictive models. The importance of a well-defined tagging scheme and consistent ITA has been well recognized and studied in the past (Poesio, 2004; Poesio and Artstein, 2005; Passonneau, 2004). There is a growing consensus that in order for these to be most useful for language understanding applications such as question answering or distillation – both of which seek to take information access technology to the next level – we need more consistent annotation of larger amounts of broad coverage data for training better automatic techniques for entity and event identification. Identification and encoding of richer knowledge – possibly linked to knowledge sources – and development of learning algorithms that would effectively incorporate them is a necessary next step towards improving the current state of the art. The computational learning community, in general, is also witnessing a move towards evaluations based on joint inference, with the two previous CoNLL tasks (Surdeanu et al., 2008; Hajič et al., 2009) devoted to joint learning of syntactic and semantic dependencies. A principle ingredient for joint learning is the presence of multiple layers of semantic information.

One fundamental question still remains, and that is – what would it take to improve the state of the art in coreference resolution that has not been attempted so far? Many different algorithms have been tried in the past 15 years, but one thing that is still lacking is a corpus comprehensively tagged on a large scale with consistent, multiple layers of semantic information. One of the many goals of the OntoNotes project[2] (Hovy et al., 2006; Weischedel et al., 2011) is to explore whether it can fill this void and help push the progress further – not only in coreference, but with the various layers of semantics that it tries to capture. As one of its layers, it has created a corpus for general anaphoric coreference that covers entities and events not limited to noun phrases or a limited set of entity types. A small portion of this corpus from the newswire and broadcast news genres (∼120k) was recently used for a SEMEVAL task (Recasens et al., 2010). As mentioned earlier, the coreference layer in OntoNotes constitutes just one part of a multi-layered, integrated annotation of shallow semantic structure in text with high inter-annotator agreement, which also provides a unique opportunity for performing joint inference over a substantial body of data.

The remainder of this paper is organized as follows. Section 2 presents an overview of the OntoNotes corpus. Section 3 describes the coreference annotation in OntoNotes. Section 4 then describes the shared task, including the data provided and the evaluation criteria. Sections 5 and 6 then describe the participating system results and analyze the approaches, and Section 7 concludes.

2 The OntoNotes Corpus

The OntoNotes project has created a corpus of large-scale, accurate, and integrated annotation of multiple levels of the shallow semantic structure in text. The idea is that this rich, integrated annotation covering many layers will allow for richer, cross-layer models enabling significantly better automatic semantic analysis. In addition to coreference, this data is also tagged with syntactic trees, high coverage verb and some noun propositions, partial verb and noun word senses, and 18 named entity types. However, such multi-layer annotations, with complex, cross-layer dependencies, demands a robust, efficient, scalable mechanism for storing them while providing efficient, convenient, integrated access to the the underlying structure. To this effect, it uses a relational database representation that captures both the inter- and intra-layer dependencies and also provides an object-oriented API for efficient, multi-tiered access to this data (Pradhan et al., 2007a). This should facilitate the creation of cross-layer features in integrated predictive models that will make use of these annotations.

Although OntoNotes is a multi-lingual resource with all layers of annotation covering three languages: English, Chinese and Arabic, for the scope of this paper, we will just look at the English portion. Over the years of the development of this corpus, there were various priorities that came into play, and therefore not all the data in the English portion is annotated with all the different layers of annotation. There is a core portion, however, which is roughly

[1] http://projects.ldc.upenn.edu/ace/data/

[2] http://www.bbn.com/nlp/ontonotes

1.3M words which has been annotated with all the layers. It comprises ∼450k words from newswire, ∼150k from magazine articles, ∼200k from broadcast news, ∼200k from broadcast conversations and ∼200k web data.

OntoNotes comprises the following layers of annotation:

- **Syntax –** A syntactic layer representing a revised Penn Treebank (Marcus et al., 1993; Babko-Malaya et al., 2006).
- **Propositions –** The proposition structure of verbs in the form of a revised PropBank(Palmer et al., 2005; Babko-Malaya et al., 2006).
- **Word Sense –** Coarse grained word senses are tagged for the most frequent polysemous verbs and nouns, in order to maximize coverage. The word sense granularity is tailored to achieve 90% inter-annotator agreement as demonstrated by Palmer et al. (2007). These senses are defined in the sense inventory files and each individual sense has been connected to multiple WordNet senses. This provides a direct access to the WordNet semantic structure for users to make use of. There is also a mapping from the word senses to the PropBank frames and to VerbNet (Kipper et al., 2000) and FrameNet (Fillmore et al., 2003).
- **Named Entities –** The corpus was tagged with a set of 18 proper named entity types that were well-defined and well-tested for inter-annotator agreement by Weischedel and Burnstein (2005).
- **Coreference –** This layer captures general anaphoric coreference that covers entities and events not limited to noun phrases or a limited set of entity types (Pradhan et al., 2007b). We will take a look at this in detail in the next section.

3 Coreference in OntoNotes

General anaphoric coreference that spans a rich set of entities and events – not restricted to a few types, as has been characteristic of most coreference data available until now – has been tagged with a high degree of consistency. Attributive coreference is tagged separately from the more common identity coreference.

Two different types of coreference are distinguished in the OntoNotes data: Identical (IDENT), and Appositive (APPOS). Appositives are treated separately because they function as attributions, as described further below. The IDENT type is used for anaphoric coreference, meaning links between pronominal, nominal, and named mentions of specific referents. It does not include mentions of generic, underspecified, or abstract entities.

Coreference is annotated for all specific entities and events. There is no limit on the semantic types of NP entities that can be considered for coreference, and in particular, coreference is not limited to ACE types.

The mentions over which IDENT coreference applies are typically pronominal, named, or definite nominal. The annotation process begins by automatically extracting all of the NP mentions from the Penn Treebank, though the annotators can also add additional mentions when appropriate. In the following two examples (and later ones), the phrases notated in bold form the links of an IDENT chain.

(1) She had **a good suggestion** and **it** was unanimously accepted by all.

(2) **Elco Industries Inc.** said **it** expects net income in the year ending June 30, 1990, to fall below a recent analyst's estimate of $ 1.65 a share. **The Rockford, Ill. maker of fasteners** also said **it** expects to post sales in the current fiscal year that are "slightly above" fiscal 1989 sales of $ 155 million.

3.1 Verbs

Verbs are added as single-word spans if they can be coreferenced with a noun phrase or with another verb. The intent is to annotate the VP, but we mark the single-word head for convenience. This includes morphologically related nominalizations (3) and noun phrases that refer to the same event, even if they are lexically distinct from the verb (4). In the following two examples, only the chains related to the *growth* event are shown.

(3) Sales of passenger cars **grew** 22%. **The strong growth** followed year-to-year increases.

(4) Japan's domestic sales of cars, trucks and buses in October **rose** 18% from a year earlier to 500,004 units, a record for the month, the Japan Automobile Dealers' Association said. The strong **growth** followed year-to-year increases of 21% in August and 12% in September.

3.2 Pronouns

All pronouns and demonstratives are linked to anything that they refer to, and pronouns in quoted speech are also marked. Expletive or pleonastic pronouns (*it, there*) are not considered for tagging, and generic *you* is not marked. In the following example, the pronoun *you* and *it* would not be marked. (In this and following examples, an asterisk (*) before a boldface phrase identifies entity/event mentions that would *not* be tagged as coreferent.)

(5) Senate majority leader Bill Frist likes to tell a story from his days as a pioneering heart surgeon back in Tennessee. A lot of times, Frist recalls, ***you'd** have a critical patient lying there waiting for a new heart, and ***you'd** want to cut, but ***you** couldn't start unless ***you** knew that the replacement heart would make ***it** to the operating room.

3.3 Generic mentions

Generic nominal mentions can be linked with referring pronouns and other definite mentions, but are not linked to other generic nominal mentions. This would allow linking of the bracketed mentions in (6) and (7), but not (8).

(6) **Officials** said **they** are tired of making the same statements.

(7) **Meetings** are most productive when **they** are held in the morning. **Those meetings**, however, generally have the worst attendance.

(8) Allergan Inc. said it received approval to sell the PhacoFlex intraocular lens, the first foldable silicone lens available for ***cataract surgery**. The lens' foldability enables it to be inserted in smaller incisions than are now possible for ***cataract surgery**.

Bare plurals, as in (6) and (7), are always considered generic. In example (9) below, there are two generic instances of *parents*. These are marked as distinct IDENT chains (with separate chains distinguished by subscripts X, Y and Z), each containing a generic and the related referring pronouns.

(9) **Parents**$_X$ should be involved with **their**$_X$ children's education at home, not in school. **They**$_X$ should see to it that **their**$_X$ kids don't play truant; **they**$_X$ should make certain that the children spend enough time doing homework; **they**$_X$ should scrutinize the report card. **Parents**$_Y$ are too likely to blame schools for the educational limitations of **their**$_Y$ children. If **parents**$_Z$ are dissatisfied with a school, **they**$_Z$ should have the option of switching to another.

In (10) below, the verb "halve" cannot be linked to "a reduction of 50%", since "a reduction" is indefinite.

(10) Argentina said it will ask creditor banks to ***halve** its foreign debt of $64 billion – the third-highest in the developing world . Argentina aspires to reach ***a reduction of 50%** in the value of its external debt.

3.4 Pre-modifiers

Proper pre-modifiers can be coreferenced, but proper nouns that are in a morphologically adjectival form are treated as adjectives, and not coreferenced. For example, adjectival forms of GPEs such as *Chinese* in "the Chinese leader", would not be linked. Thus we could coreference *United States* in "the United States policy" with another referent, but not *American* "the American policy." GPEs and Nationality acronyms (e.g. *U.S.S.R.* or *U.S.*). are also considered adjectival. Pre-modifier acronyms can be coreferenced unless they refer to a nationality. Thus in the examples below, *FBI* can be coreferenced to other mentions, but *U.S.* cannot.

(11) **FBI** spokesman

(12) ***U.S.** spokesman

Dates and monetary amounts can be considered part of a coreference chain even when they occur as pre-modifiers.

(13) The current account deficit on France's balance of payments narrowed to 1.48 billion French francs ($236.8 million) in August from a revised 2.1 billion francs in **July**, the Finance Ministry said. Previously, the **July** figure was estimated at a deficit of 613 million francs.

(14) The company's **$150** offer was unexpected. The firm balked at **the price**.

3.5 Copular verbs

Attributes signaled by copular structures are not marked; these are attributes of the referent they modify, and their relationship to that referent will be captured through word sense and propositional argument tagging.

(15) **John**$_X$ is a linguist. **People**$_Y$ are nervous around **John**$_X$, because **he**$_X$ always corrects **their**$_Y$ grammar.

Copular (or 'linking') verbs are those verbs that function as a copula and are followed by a subject complement. Some common copular verbs are: *be, appear, feel, look, seem, remain, stay, become, end up, get*. Subject complements following such verbs are considered attributes, and not linked. Since *Called* is copular, neither IDENT nor APPOS coreference is marked in the following case.

(16) Called Otto's Original Oat Bran Beer, the brew costs about $12.75 a case.

3.6 Small clauses

Like copulas, small clause constructions are not marked. The following example is treated as if the copula were present ("John considers Fred to be an idiot"):

(17) John considers ***Fred *an idiot**.

3.7 Temporal expressions

Temporal expressions such as the following are linked:

(18) John spent **three years** in jail. In **that time**...

Deictic expressions such as *now, then, today, tomorrow, yesterday,* etc. can be linked, as well as other temporal expressions that are relative to the time of the writing of the article, and which may therefore require knowledge of the time of the writing to resolve the coreference. Annotators were allowed to use knowledge from outside the text in resolving these cases. In the following example, *the end of this period* and *that time* can be coreferenced, as can *this period* and *from three years to seven years*.

(19) The limit could range **from three years to seven years**$_X$, depending on the composition of the management team and the nature of its strategic plan. At **(the end of (this period)**$_X$**)**$_Y$, the poison pill would be eliminated automatically, unless a new poison pill were approved by the then-current shareholders, who would have an opportunity to evaluate the corporation's strategy and management team at **that time**$_Y$.

In multi-date temporal expressions, embedded dates are not separately connected to to other mentions of that date. For example in *Nov. 2, 1999*, *Nov.* would not be linked to another instance of *November* later in the text.

3.8 Appositives

Because they logically represent attributions, appositives are tagged separately from Identity coreference. They consist of a head, or referent (a noun phrase that points to a specific object/concept in the world), and one or more attributes of that referent. An appositive construction contains a noun phrase that modifies an immediately-adjacent noun phrase (separated only by a comma, colon, dash, or parenthesis). It often serves to rename or further define the first mention. Marking appositive constructions allows us to capture the attributed property even though there is no explicit copula.

(20) **John**$_{\text{head}}$, **a linguist**$_{\text{attribute}}$

The head of each appositive construction is distinguished from the attribute according to the following heuristic specificity scale, in a decreasing order from top to bottom:

Type	Example
Proper noun	John
Pronoun	He
Definite NP	the man
Indefinite specific NP	a man I know
Non-specific NP	man

This leads to the following cases:

(21) **John**$_{\text{head}}$, **a linguist**$_{\text{attribute}}$

(22) **A famous linguist**$_{\text{attribute}}$, **he**$_{\text{head}}$ studied at ...

(23) **a principal of the firm**$_{\text{attribute}}$, **J. Smith**$_{\text{head}}$

In cases where the two members of the appositive are equivalent in specificity, the left-most member of the appositive is marked as the head/referent. Definite NPs include NPs with a definite marker (*the*) as well as NPs with a possessive adjective (his). Thus the first element is the head in all of the following cases:

(24) The chairman, the man who never gives up

(25) The sheriff, his friend

(26) His friend, the sheriff

In the specificity scale, specific names of diseases and technologies are classified as proper names, whether they are capitalized or not.

(27) A dangerous bacteria, bacillium, is found

Type	Description
Annotator Error	An annotator error. This is a catch-all category for cases of errors that do not fit in the other categories.
Genuine Ambiguity	This is just genuinely ambiguous. Often the case with pronouns that have no clear antecedent (especially this & that)
Generics	One person thought this was a generic mention, and the other person didn't
Guidelines	The guidelines need to be clear about this example
Callisto Layout	Something to do with the usage/design of Callisto
Referents	Each annotator thought this was referring to two completely different things
Possessives	One person did not mark this possessive
Verb	One person did not mark this verb
Pre Modifiers	One person did not mark this Pre Modifier
Appositive	One person did not mark this appositive
Extent	Both people marked the same entity, but one person's mention was longer
Copula	Disagreement arose because this mention is part of a copular structure a) Either each annotator marked a different half of the copula b) Or one annotator unnecessarily marked both

Figure 1: Description of various disagreement types

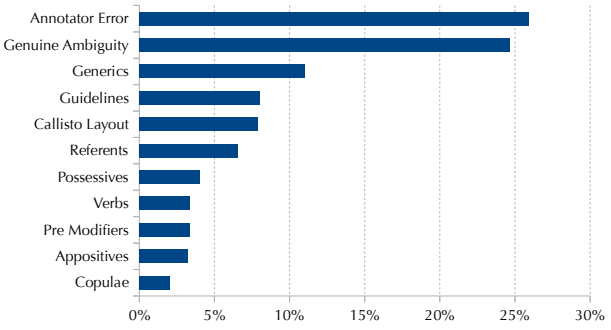

Figure 2: The distribution of disagreements across the various types in Table 1

When the entity to which an appositive refers is also mentioned elsewhere, only the single span containing the entire appositive construction is included in the larger IDENT chain. None of the nested NP spans are linked. In the example below, the entire span can be linked to later mentions to *Richard Godown*. The sub-spans are not included separately in the IDENT chain.

(28) **Richard Godown, president of the Industrial Biotechnology Association**

Ages are tagged as attributes (as if they were ellipses of, for example, *a 42-year-old*):

(29) **Mr.Smith**$_{\text{head}}$, **42**$_{\text{attribute}}$,

3.9 Special Issues

In addition to the ones above, there are some special cases such as:

- No coreference is marked between an organization and its members.

Genre	ANN1-ANN2	ANN1-ADJ	ANN2-ADJ
Newswire	80.9	85.2	88.3
Broadcast News	78.6	83.5	89.4
Broadcast Conversation	86.7	91.6	93.7
Magazine	78.4	83.2	88.8
Web	85.9	92.2	91.2

Table 1: Inter Annotator and Adjudicator agreement for the Coreference Layer in OntoNotes measured in terms of the MUC score.

- GPEs are linked to references to their governments, even when the references are nested NPs, or the modifier and head of a single NP.

3.10 Annotator Agreement and Analysis

Table 1 shows the inter-annotator and annotator-adjudicator agreement on all the genres of OntoNotes. We also analyzed about 15K disagreements in various parts of the data, and grouped them into one of the categories shown in Figure 1. Figure 2 shows the distribution of these different types that were found in that sample. It can be

seen that genuine ambiguity and annotator error are the biggest contributors – the latter of which is usually captured during adjudication, thus showing the increased agreement between the adjudicated version and the individual annotator version.

4 CoNLL-2011 Coreference Task

This section describes the CoNLL-2011 Coreference task, including its *closed* and *open* track versions, and characterizes the data used for the task and how it was prepared.

4.1 Why a Coreference Task?

Despite close to a two-decade history of evaluations on coreference tasks, variation in the evaluation criteria and in the training data used have made it difficult for researchers to be clear about the state of the art or to determine which particular areas require further attention. There are many different parameters involved in defining a coreference task. Looking at various numbers reported in literature can greatly affect the perceived difficulty of the task. It can seem to be a very hard problem (Soon et al., 2001) or one that is somewhat easier (Culotta et al., 2007). Given the space constraints, we refer the reader to Stoyanov et al. (2009) for a detailed treatment of the issue.

Limitations in the size and scope of the available datasets have also constrained research progress. The MUC and ACE corpora are the two that have been used most for reporting comparative results, but they differ in the types of entities and coreference annotated. The ACE corpus is also one that evolved over a period of almost five years, with different incarnations of the task definition and different corpus cross-sections on which performance numbers have been reported, making it hard to untangle and interpret the results.

The availability of the OntoNotes data offered an opportunity to define a coreference task based on a larger, more broad-coverage corpus. We have tried to design the task so that it not only can support the current evaluation, but also can provide an ongoing resource for comparing different coreference algorithms and approaches.

4.2 Task Description

The CoNLL-2011 shared task was based on the English portion of the OntoNotes 4.0 data. The task was to automatically identify mentions of entities and events in text and to link the coreferring mentions together to form entity/event chains. The target coreference decisions could be made using automatically predicted information on the other structural layers including the parses, semantic roles, word senses, and named entities.

As is customary for CoNLL tasks, there were two tracks, *closed* and *open*. For the *closed* track, systems were limited to using the distributed resources, in order to allow a fair comparison of algorithm performance, while the *open* track allowed for almost unrestricted use of external resources in addition to the provided data.

4.2.1 Closed Track

In the *closed* track, systems were limited to the provided data, plus the use of *two pre-specified external resources*: i) WordNet and ii) a pre-computed number and gender table by Bergsma and Lin (2006).

For the training and test data, in addition to the underlying text, *predicted* versions of all the supplementary layers of annotation were provided, where those predictions were derived using off-the-shelf tools (parsers, semantic role labelers, named entity taggers, etc.) as described in Section 4.4.2. For the training data, however, in addition to predicted values for the other layers, we also provided manual *gold-standard* annotations for all the layers. Participants were allowed to use either the gold-standard or predicted annotation for training their systems. They were also free to use the gold-standard data to train their own models for the various layers of annotation, if they judged that those would either provide more accurate predictions or alternative predictions for use as multiple views, or wished to use a lattice of predictions.

More so than previous CoNLL tasks, coreference predictions depend on world knowledge, and many state-of-the-art systems use information from external resources such as WordNet, which can add a layer that helps the system to recognize semantic connections between the various lexicalized mentions in the text. Therefore, the use of WordNet was allowed, even for the closed track. Since word senses in OntoNotes are predominantly[3] coarse-grained groupings of WordNet senses, systems could also map from the predicted or gold-standard word senses provided to the sets of underlying WordNet senses. Another significant piece of knowledge that is particularly useful for coreference but that is not available in the layers of OntoNotes is that of *number* and *gender*. There are many different

[3]There are a few instances of novel senses introduced in OntoNotes which were not present in WordNet, and so lack a mapping back to the WordNet senses

ways of predicting these values, with differing accuracies, so in order to ensure that participants in the *closed* track were working from the same data, thus allowing clearer algorithmic comparisons, we specified a particular table of number and gender predictions generated by Bergsma and Lin (2006), for use during both training and testing.

Following the recent CoNLL tradition, participants were allowed to use both the training and the development data for training the final model.

4.2.2 Open Track

In addition to resources available in the *closed* track, the *open* track, systems were allowed to use external resources such as Wikipedia, gazetteers etc. This track is mainly to get an idea of a performance ceiling on the task at the cost of not getting a comparison across all systems. Another advantage of the *open* track is that it might reduce the barriers to participation by allowing participants to field existing research systems that already depend on external resources – especially if there were hard dependencies on these resources. They can participate in the task with minimal or no modification to their existing system.

4.3 Coreference Task Data

Since there are no previously reported numbers on the full version of OntoNotes, we had to create a train/development/test partition. The only portion of OntoNotes that has a previously determined, widely used, standard split is the WSJ portion of the newswire data. For that subcorpus, we maintained the same partition. For all the other portions we created stratified training, development and test partitions over all the sources in OntoNotes using the procedure shown in Algorithm 1. The list of training, development and test document IDs can be found on the task webpage.[4]

4.4 Data Preparation

This section gives details of the different annotation layers including the automatic models that were used to predict them, and describes the formats in which the data were provided to the participants.

4.4.1 Manual Annotation *Gold* Layers

We will take a look at the manually annotated, or *gold* layers of information that were made available for the training data.

[4] http://conll.bbn.com/download/conll-train.id
http://conll.bbn.com/download/conll-dev.id
http://conll.bbn.com/download/conll-test.id

Algorithm 1 Procedure used to create OntoNotes training, development and test partitions.

Procedure: GENERATE_PARTITIONS(ONTONOTES) **returns** TRAIN, DEV, TEST

```
TRAIN ← ∅
DEV ← ∅
TEST ← ∅
for all SOURCE ∈ ONTONOTES do
    if SOURCE = WALL STREET JOURNAL then
        TRAIN ← TRAIN ∪ SECTIONS 02 – 21
        DEV ← DEV ∪ SECTIONS 00, 01, 22, 24
        TEST ← TEST ∪ SECTION 23
    else
        if Number of files in SOURCE ≥ 10 then
            TRAIN ← TRAIN ∪ FILE IDS ending in 1 – 8
            DEV ← DEV ∪ FILE IDS ending in 0
            TEST ← TEST ∪ FILE IDS ending in 9
        else
            DEV ← DEV ∪ FILE IDS ending in 0
            TEST ← TEST ∪ FILE ID ending in the highest number
            TRAIN ← TRAIN ∪ Remaining FILE IDS for the
            SOURCE
        end if
    end if
end for
return TRAIN, DEV, TEST
```

Coreference The manual coreference annotation is stored as chains of linked mentions connecting multiple mentions of the same entity. Coreference is the only document-level phenomenon in OntoNotes, and the complexity of annotation increases non-linearly with the length of a document. Unfortunately, some of the documents – especially ones in the broadcast conversation, weblogs, and telephone conversation genre – are very long which prohibited us from efficiently annotating them in entirety. These had to be split into smaller parts. We conducted a few passes to join some adjacent parts, but since some documents had as many as 17 parts, there are still multi-part documents in the corpus. Since the coreference chains are coherent only within each of these document parts, for this task, each such part is treated as a separate document. Another thing to note is that there were some cases of sub-token annotation in the corpus owing to the fact that tokens were not split at hyphens. Cases such as pro-WalMart had the sub-span WalMart linked with another instance of the same. The recent Treebank revision which split tokens at *most* hyphens, made a majority of these sub-token annotations go away. There were still some residual sub-token annotations. Since subtoken annotations cannot be represented in the CoNLL format, and they were a very small quantity – much less than even half a percent – we decided to ignore them.

For various reasons, not all the documents in OntoNotes have been annotated with all the differ-

Corpora	Words				Documents			
	Total	Train	Dev	Test	Total	Train	Dev	Test
MUC-6	25K	12K	13K		60	30	30	
MUC-7	40K	19K	21K		67	30	37	
ACE (2000-2004)	1M	775K	235K		-	-	-	
OntoNotes[5]	1.3M	1M	136K	142K	2,083 (2,999)	1,674 (2,374)	202 (303)	207 (322)

Table 2: Number of documents in the OntoNotes data, and some comparison with the MUC and ACE data sets. The numbers in parenthesis for the OntoNotes corpus indicate the total number of *parts* that correspond to the documents. Each part was considered a separate document for evaluation purposes.

Syntactic category	Train		Development		Test	
	Count	%	Count	%	Count	%
NP	60,345	59.71	8,463	59.31	8,629	53.09
PRP	25,472	25.21	3,535	24.78	5,012	30.84
PRP$	8,889	8.80	1,208	8.47	1,466	9.02
NNP	2,643	2.62	468	3.28	475	2.92
NML	900	0.89	151	1.06	118	0.73
Vx	1,915	1.89	317	2.22	314	1.93
Other	893	0.88	126	0.88	239	1.47
Overall	101,057	100.00	14,268	100.00	16,253	100.00

Table 3: Distribution of mentions in the data by their syntactic category.

	Train	Development	Test
Entities/Chains	26,612	3,752	3,926
Links	74,652	10,539	12,365
Mentions	101,264	14,291	16,291

Table 4: Number of entities, links and mentions in the OntoNotes 4.0 data.

ent layers of annotation, with full coverage.[6] There is a core portion, however, which is roughly 1.3M words which has been annotated with all the layers. This is the portion that we used for the shared task.

The number of documents in the corpus for this task, for each of the different genres, are shown in Table 2. Tables 3 and 4 shows the distribution of mentions by the syntactic categories, and the counts of entities, links and mentions in the corpus respectively. All of this data has been Treebanked and PropBanked either as part of the OntoNotes effort or some preceding effort.

For comparison purposes, Table 2 also lists the number of documents in the MUC-6, MUC-7, and ACE (2000-2004) corpora. The MUC-6 data was taken from the Wall Street Journal, whereas the MUC-7 data was from the New York Times. The ACE data spanned many different genres similar to the ones in OntoNotes.

Parse Trees This represents the syntactic layer that is a revised version of the Penn Treebank. For purposes of this task, traces were removed from the syntactic trees, since the CoNLL-style data format, being indexed by tokens, does not provide any good means of conveying that information. Function tags were also removed, since the parsers that we used for the predicted syntax layer did not provide them. One thing that needs to be dealt with in conversational data is the presence of disfluencies (restarts, etc.). In the original OntoNotes parses, these are marked using a special EDITED[7] phrase tag – as was the case for the Switchboard Treebank. Given the frequency of disfluencies and the performance with which one can identify them automatically,[8] a probable processing pipeline would filter them out before parsing. Since we did not have a readily available tagger for tagging disfluencies, we decided to remove them using oracle information available in the Treebank.

Propositions The propositions in OntoNotes constitute PropBank semantic roles. Most of the verb predicates in the corpus have been annotated with their arguments. Recent enhancements to the PropBank to make it synchronize better with the Treebank (Babko-Malaya et al., 2006) have enhanced the information in the proposition by the addition of two types of LINKs that represent pragmatic coreference (LINK-PCR) and selectional preferences (LINK-SLC). More details can be found in the addendum to the PropBank guidelines[9] in the OntoNotes 4.0 re-

[6]Given the nature of word sense annotation, and changes in project priorities, we could not annotate all the low frequency verbs and nouns in the corpus. Furthermore, PropBank annotation currently only covers verb predicates.

[7]There is another phrase type – EMBED in the telephone conversation genre which is similar to the EDITED phrase type, and sometimes identifies insertions, but sometimes contains logical continuation of phrases, so we decided not to remove that from the data.

[8]A study by Charniak and Johnson (2001) shows that one can identify and remove edits from transcribed conversational speech with an F-score of about 78, with roughly 95 Precision and 67 recall.

[9]doc/propbank/english-propbank.pdf

lease. Since the community is not used to this representation which relies heavily on the trace structure in the Treebank which we are excluding, we decided to *unfold* the LINKs back to their original representation as in the Release 1.0 of the Proposition Bank. This functionality is part of the OntoNotes DB Tool.[10]

Word Sense Gold word sense annotation was supplied using sense numbers as specified in the OntoNotes list of senses for each lemma.[11] The sense inventories that were provided in the OntoNotes 4.0 release were not all mapped to the latest version 3.0 of WordNet, so we provided a revised version of the sense inventories, containing mapping to WordNet 3.0, on the task page for the participants.

Named Entities Named Entities in OntoNotes data are specified using a catalog of 18 Name types.

Other Layers Discourse plays a vital role in coreference resolution. In the case of broadcast conversation, or telephone conversation data, it partially manifests in the form of speakers of a given utterance, whereas in weblogs or newsgroups it does so as the writer, or commenter of a particular article or thread. This information provides an important clue for correctly linking anaphoric pronouns with the right antecedents. This information could be automatically deduced, but since it would add additional complexity to the already complex task, we decided to provide oracle information of this metadata both during training and testing. In other words, speaker and author identification was not treated as an annotation layer that needed to be predicted. This information was provided in the form of another column in the `.conll` table. There were some cases of interruptions and interjections that ideally would associate parts of a sentence to two different speakers, but since the frequency of this was quite small, we decided to make an assumption of one speaker/writer per sentence.

4.4.2 Predicted Annotation Layers

The predicted annotation layers were derived using automatic models trained using cross-validation on other portions of OntoNotes data. As mentioned earlier, there are some portions of the OntoNotes corpus that have not been annotated for coreference but that have been annotated for other layers. For training models for each of the layers, where feasible, we used all the data that we could for that layer from the training portion of the entire OntoNotes release.

Senses	Lemmas
1	1,506
2	1,046
> 2	1,016

Table 6: Word sense polysemy over verb and noun lemmas in OntoNotes

Parse Trees Predicted parse trees were produced using the Charniak parser (Charniak and Johnson, 2005).[12] Some additional tag types used in the OntoNotes trees were added to the parser's tagset, including the NML tag that has recently been added to capture internal NP structure, and the rules used to determine head words were appropriately extended. The parser was then re-trained on the training portion of the release 4.0 data using 10-fold cross-validation. Table 5 shows the performance of the re-trained Charniak parser on the CoNLL-2011 test set. We did not get a chance to re-train the re-ranker, and since the stock re-ranker crashes when run on n-best parses containing NMLs, because it has not seen that tag in training, we could not make use of it.

Word Sense We trained a word sense tagger using a SVM classifier and contextual word and part of speech features on all the training portion of the OntoNotes data. The OntoNotes 4.0 corpus comprises a total of 14,662 sense definitions across 4877 verb and noun lemmas[13]. The distribution of senses per lemma is as shown in Table 6. Table 7 shows the performance of this classifier over *both the verbs and nouns* in the CoNLL-2011 test set. Again this performance is not directly comparable to any reported in the literature before, and it seems lower then performances reported on previous versions of OntoNotes because this is over all the genres of OntoNotes, and aggregated over both verbs and nouns in the CoNLL-2011 test set.

Propositions To predict propositional structure, ASSERT[14] (Pradhan et al., 2005) was used, retrained also on all the training portion of the release

[10] http://cemantix.org/ontonotes.html

[11] It should be noted that word sense annotation in OntoNotes is note complete, so only some of the verbs and nouns have word sense tags specified.

[12] http://bllip.cs.brown.edu/download/reranking-parserAug06.tar.gz

[13] The number of lemmas in Table 6 do not add up to this number because not all of them have examples in the training data, where the total number of instantiated senses amounts to 7933.

[14] http://cemantix.org/assert.html

	All Sentences					Sentence len < 40			
	N	POS	R	P	F	N	R	P	F
Broadcast Conversation (BC)	2,194	95.93	84.30	84.46	84.38	2124	85.83	85.97	85.90
Broadcast News (BN)	1,344	96.50	84.19	84.28	84.24	1278	85.93	86.04	85.98
Magazine (MZ)	780	95.14	87.11	87.46	87.28	736	87.71	88.04	87.87
Newswire (NW)	2,273	96.95	87.05	87.45	87.25	2082	88.95	89.27	89.11
Telephone Conversation (TC)	1,366	93.52	79.73	80.83	80.28	1359	79.88	80.98	80.43
Weblogs and Newsgroups (WB)	1,658	94.67	83.32	83.20	83.26	1566	85.14	85.07	85.11
Overall	9,615	96.03	85.25	85.43	85.34	9145	86.86	87.02	86.94

Table 5: Parser performance on the CoNLL-2011 test set

	Frameset Accuracy	Total Sentences	Total Propositions	% Perfect Propositions	Argument ID + Class		
					P	R	F
Broadcast Conversation (BC)	0.92	2,037	5,021	52.18	82.55	64.84	72.63
Broadcast News (BN)	0.91	1,252	3,310	53.66	81.64	64.46	72.04
Magazine (MZ)	0.89	780	2,373	47.16	79.98	61.66	69.64
Newswire (NW)	0.93	1,898	4,758	39.72	80.53	62.68	70.49
Weblogs and Newsgroups (WB)	0.92	929	2,174	39.19	81.01	60.65	69.37
Overall	0.91	6,896	17,636	46.82	81.28	63.17	71.09

Table 8: Performance on the propositions and framesets in the CoNLL-2011 test set.

	Accuracy
Broadcast Conversation (BC)	0.70
Broadcast News (BN)	0.68
Magazine (MZ)	0.60
Newswire (NW)	0.62
Weblogs and Newsgroups (WB)	0.63
Overall	0.65

Table 7: Word sense performance over both verbs and nouns in the CoNLL-2011 test set

4.0 data. Given time constraints, we had to perform two modifications: i) Instead of a single model that predicts all arguments including NULL arguments, we had to use the two-stage mode where the NULL arguments are first filtered out and the remaining NON-NULL arguments are classified into one of the argument types, and ii) The argument identification module used an ensemble of ten classifiers – each trained on a tenth of the training data and performed an unweighted voting among them. This should still give a close to state of the art performance given that the argument identification performance tends to start to be asymptotic around 10k training instances. At first glance, the performance on the newswire genre is much lower than what has been reported for WSJ Section 23. This could be attributed to two factors: i) the fact that we had to compromise on the training method, but more importantly because ii) the newswire in OntoNotes not only contains WSJ data, but also Xinhua news. One could try to verify using just the WSJ portion of the data, but it would be hard as it is not only a subset of the documents that the performance has been reported on previously, but also the annotation has been significantly revised; it includes propositions for *be* verbs missing from the original PropBank, and the training data is a subset of the original data as well. Table 8 shows the detailed performance numbers.

In addition to automatically predicting the arguments, we also trained a classifier to tag PropBank frameset IDs in the data using the same word sense module as mentioned earlier. OntoNotes 4.0 contains a total of 7337 framesets across 5433 verb lemmas.[15] An overwhelming number of them are monosemous, but the more frequent verbs tend to be polysemous. Table 9 gives the distribution of number of framesets per lemma in the PropBank layer of the OntoNotes 4.0 data.

During automatic processing of the data, we tagged all the tokens that were tagged with a part of speech VBx. This means that there would be cases where the wrong token would be tagged with propositions. The CoNLL-2005 scorer was used to generate the scores.

Named Entities BBN's IdentiFinder™system was used to predict the named entities. Given the

[15]The number of lemmas in Table 9 do not add up to this number because not all of them have examples in the training data, where the total number of instantiated senses amounts to 4229.

Framesets	Lemmas
1	2,722
2	321
> 2	181

Table 9: Frameset polysemy across lemmas

	Overall	BC	BN	MZ	NW	TC	WB
	F	F	F	F	F	F	F
ALL Named Entities	71.8	64.8	72.2	61.5	84.3	39.5	55.2
Cardinal	68.7	51.8	71.1	66.1	82.8	34.0	68.7
Date	76.1	63.7	77.9	66.7	83.7	60.5	56.0
Event	27.6	00.0	34.8	30.8	47.6	-	13.3
Facility	41.9	55.0	16.7	23.1	66.7	00.0	22.9
GPE	87.9	87.5	90.3	73.7	92.9	65.9	88.7
Language	41.2	-	50.0	50.0	00.0	20.0	75.0
Law	63.0	00.0	85.7	00.0	67.9	00.0	50.0
Location	58.4	59.1	59.6	53.3	68.0	00.0	23.5
Money	74.6	16.7	66.7	73.2	79.4	30.8	61.5
NORP	00.0	00.0	00.0	00.0	00.0	00.0	00.0
Ordinal	73.4	73.8	73.4	78.1	78.4	88.9	37.0
Organization	71.0	57.8	67.1	52.9	86.9	21.2	32.1
Percent	71.2	88.9	76.9	69.6	92.1	01.2	71.6
Person	79.6	78.9	87.7	66.7	91.6	65.1	64.8
Product	46.9	00.0	43.8	00.0	81.8	00.0	00.0
Quantity	47.5	25.3	58.3	61.1	71.9	00.0	22.2
Time	58.6	56.9	64.1	42.9	80.0	23.8	51.7
Work of Art	41.9	26.9	37.1	16.0	77.9	00.0	05.6

Table 10: Named Entity performance on the CoNLL-2011 test set

time constraints, we could not re-train it on the OntoNotes data and so an existing, pre-trained model was used, therefore the results are not a good indicator of the model's best performance. The pre-trained model had also used a somewhat different catalog of name types, which did not include the OntoNotes NORP type (for nationalities, organizations, religions, and political parties), so that category was never predicted. Table 10 shows the overall performance of the tagger on the CoNLL-2011 test set, as well as the performance broken down by individual name types. IdentiFinder performance has been reported to be in the low 90's on WSJ test set.

Other Layers As noted above, systems were allowed to make use of gender and number predictions for NPs using the table from Bergsma and Lin (Bergsma and Lin, 2006).

4.4.3 Data Format

In order to organize the multiple, rich layers of annotation, the OntoNotes project has created a database representation for the raw annotation layers along with a Python API to manipulate them (Pradhan et al., 2007a). In the OntoNotes distribution the data is organized as one file per layer, per document. The API requires a certain hierarchical structure with documents at the leaves inside a hierarchy of language, genre, source and section. It comes with various ways of cleanly querying and manipulating the data and allows convenient access to the sense inventory and propbank frame files instead of having to interpret the raw `.xml` versions. However, maintaining format consistency with earlier CoNLL tasks was deemed convenient for sites that already had tools configured to deal with that format. Therefore, in order to distribute the data so that one could make the best of both worlds, we created a new file type called `.conll` which logically served as another layer in addition to the `.parse`, `.prop`, `.name` and `.coref` layers. Each `.conll` file contained a merged representation of all the OntoNotes layers in the CoNLL-style tabular format with one line per token, and with multiple columns for each token specifying the input annotation layers relevant to that token, with the final column specifying the target coreference layer. Because OntoNotes is not authorized to distribute the underlying text, and many of the layers contain inline annotation, we had to provide a skeletal form (`.skel` of the `.conll` file which was essentially the `.conll` file, but with the word column replaced with a dummy string. We provided an assembly script that participants could use to create a `.conll` file taking as input the `.skel` file and the top-level directory of the OntoNotes distribution that they had separately downloaded from the LDC[16] Once the `.conll` file is created, it can be used to create the individual layers such as `.parse`, `.name`, `.coref` etc. using another set of scripts. Since the propositions and word sense layers are inherently standoff annotation, they were provided as is, and did not require that extra merging step. One thing thing that made this data creation process a bit tricky was the fact that we had dissected some of the trees for the conversation data to remove the EDITED phrases. Table 11 describes the data provided in each of the column of the `.conll` format. Figure 3 shows a sample from a `.conll` file.

4.5 Evaluation

This section describes the evaluation criteria used. Unlike for propositions, word sense and named entities, where it is simply a matter of counting the correct answers, or for parsing, where there are several established metrics, evaluating the accuracy of coreference continues to be contentious. Various al-

[16]OntoNotes is deeply grateful to the Linguistic Data Consortium for making the source data freely available to the task participants.

Column	Type	Description
1	Document ID	This is a variation on the document filename
2	Part number	Some files are divided into multiple parts numbered as 000, 001, 002, ... etc.
3	Word number	This is the word index in the sentence
4	Word	The word itself
5	Part of Speech	Part of Speech of the word
6	Parse bit	This is the bracketed structure broken before the first open parenthesis in the parse, and the word/part-of-speech leaf replaced with a *. The full parse can be created by substituting the asterix with the `([pos] [word])` string (or leaf) and concatenating the items in the rows of that column.
7	Predicate lemma	The predicate lemma is mentioned for the rows for which we have semantic role information. All other rows are marked with a –
8	Predicate Frameset ID	This is the PropBank frameset ID of the predicate in Column 7.
9	Word sense	This is the word sense of the word in Column 3.
10	Speaker/Author	This is the speaker or author name where available. Mostly in Broadcast Conversation and Web Log data.
11	Named Entities	These columns identifies the spans representing various named entities.
12:N	Predicate Arguments	There is one column each of predicate argument structure information for the predicate mentioned in Column 7.
N	Coreference	Coreference chain information encoded in a parenthesis structure.

Table 11: Format of the `.conll` file used on the shared task

```
#begin document (nw/wsj/07/wsj_0771); part 000
...
...
nw/wsj/07/wsj_0771 0   0         ``   ``  (TOP(S(S*           -    -  -   -         *       *     (ARG1*          *          *       -
nw/wsj/07/wsj_0771 0   1 Vandenberg  NNP       (NP*           -    -  -   -  (PERSON)  (ARG1*          *          *          * (8|(0)
nw/wsj/07/wsj_0771 0   2        and   CC          *           -    -  -   -         *       *          *          *          *       -
nw/wsj/07/wsj_0771 0   3    Rayburn  NNP         *)           -    -  -   -  (PERSON)       *)         *          *          *(23)|8)
nw/wsj/07/wsj_0771 0   4        are  VBP       (VP*          be   01  1   -         *    (V*)          *          *          *       -
nw/wsj/07/wsj_0771 0   5     heroes  NNS    (NP(NP*)          -    -  -   -         *  (ARG2*          *          *          *       -
nw/wsj/07/wsj_0771 0   6         of   IN       (PP*           -    -  -   -         *       *          *          *          *       -
nw/wsj/07/wsj_0771 0   7       mine   NN    (NP*))))          -    -  5   -         *      *)          *          *          *    (15)
nw/wsj/07/wsj_0771 0   8          ,    ,         *           -    -  -   -         *       *          *          *          *       -
nw/wsj/07/wsj_0771 0   9         ,,   ,,         *)          -    -  -   -         *       *         *)          *          *       -
nw/wsj/07/wsj_0771 0  10        Mr.  NNP       (NP*           -    -  -   -         *       *     (ARG0*     (ARG0*          *    (15
nw/wsj/07/wsj_0771 0  11      Boren  NNP         *)           -    -  -   -  (PERSON)       *         *)         *)          *    15)
nw/wsj/07/wsj_0771 0  12       says  VBZ       (VP*        say   01  1   -         *       *       (V*)         *          *       -
nw/wsj/07/wsj_0771 0  13          ,    ,         *           -    -  -   -         *       *         *          *          *       -
nw/wsj/07/wsj_0771 0  14  referring  VBG     (S(VP*       refer   01  2   -         *       * (ARGM-ADV*       (V*)          *       -
nw/wsj/07/wsj_0771 0  15         as   RB     (ADVP*          -    -  -   -         *       *         * (ARGM-DIS*          *       -
nw/wsj/07/wsj_0771 0  16       well   RB         *)          -    -  -   -         *       *         *         *)          *       -
nw/wsj/07/wsj_0771 0  17         to   IN       (PP*          -    -  -   -         *       *         *     (ARG1*          *       -
nw/wsj/07/wsj_0771 0  18        Sam  NNP    (NP(NP*          -    -  -   -  (PERSON*       *         *          *          *    (23
nw/wsj/07/wsj_0771 0  19    Rayburn  NNP         *)          -    -  -   -        *)       *         *          *          *       -
nw/wsj/07/wsj_0771 0  20          ,    ,         *           -    -  -   -         *       *         *          *          *       -
nw/wsj/07/wsj_0771 0  21        the   DT    (NP(NP*          -    -  -   -         *       *         *          *     (ARG0*       -
nw/wsj/07/wsj_0771 0  22 Democratic   JJ         *           -    -  -   -    (NORP)       *         *          *          *       -
nw/wsj/07/wsj_0771 0  23      House  NNP         *           -    -  -   -     (ORG)       *         *          *          *       -
nw/wsj/07/wsj_0771 0  24    speaker   NN         *)          -    -  -   -         *       *         *          *         *)       -
nw/wsj/07/wsj_0771 0  25        who   WP  (SBAR(WHNP*)        -    -  -   -         *       *         *          * (R-ARG0*)       -
nw/wsj/07/wsj_0771 0  26 cooperated  VBD     (S(VP*  cooperate   01  1   -         *       *         *          *       (V*)       -
nw/wsj/07/wsj_0771 0  27       with   IN       (PP*          -    -  -   -         *       *         *          *     (ARG1*       -
nw/wsj/07/wsj_0771 0  28  President  NNP       (NP*          -    -  -   -         *       *         *          *          *       -
nw/wsj/07/wsj_0771 0  29 Eisenhower  NNP *))))))))))        -    -  -   -  (PERSON)       *         *)         *)         *)    23)
nw/wsj/07/wsj_0771 0  30          .    .         *))         -    -  -   -         *       *         *          *          *       -

nw/wsj/07/wsj_0771 0   0         ``   ``    (TOP(S*          -    -  -   -         *       *         *          -
nw/wsj/07/wsj_0771 0   1       They  PRP      (NP*)          -    -  -   -         * (ARG0*)         *        (8)
nw/wsj/07/wsj_0771 0   2    allowed  VBD       (VP*       allow   01  1   -         *    (V*)         *          -
nw/wsj/07/wsj_0771 0   3       this   DT     (S(NP*          -    -  -   -         *  (ARG1*     (ARG1*        (6
nw/wsj/07/wsj_0771 0   4    country   NN         *)          -    -  3   -         *       *         *)        6)
nw/wsj/07/wsj_0771 0   5         to   TO       (VP*          -    -  -   -         *       *         *          -
nw/wsj/07/wsj_0771 0   6         be   VB       (VP*         be   01  1   -         *       *       (V*)       (16)
nw/wsj/07/wsj_0771 0   7   credible   JJ (ADJP*)))))         -    -  -   -         *      *)    (ARG2*)          -
nw/wsj/07/wsj_0771 0   8          .    .        *))          -    -  -   -         *       *         *          -

#end document
```

Figure 3: Sample portion of the `.conll` file.

ternative metrics have been proposed, as mentioned below, which weight different features of a proposed coreference pattern differently. The choice is not clear in part because the value of a particular set of coreference predictions is integrally tied to the consuming application.

A further issue in defining a coreference metric concerns the granularity of the mentions, and how closely the predicted mentions are required to match those in the gold standard for a coreference prediction to be counted as correct.

Our evaluation criterion was in part driven by the OntoNotes data structures. OntoNotes coreference distinguishes between identity coreference and appositive coreference, treating the latter separately because it is already captured explicitly by other layers of the OntoNotes annotation. Thus we evaluated systems only on the identity coreference task, which links all categories of entities and events together into equivalent classes.

The situation with mentions for OntoNotes is also different than it was for MUC or ACE. OntoNotes data does not explicitly identify the minimum extents of an entity mention, but it does include hand-tagged syntactic parses. Thus for the official evaluation, we decided to use the exact spans of mentions for determining correctness. The NP boundaries for the test data were pre-extracted from the hand-tagged Treebank for annotation, and events triggered by verb phrases were tagged using the verbs themselves. This choice means that scores for the CoNLL-2011 coreference task are likely to be lower than for coref evaluations based on MUC, where the mention spans are specified in the input,[17] or those based on ACE data, where an approximate match is often allowed based on the specified head of the NP mention.

4.5.1 Metrics

As noted above, the choice of an evaluation metric for coreference has been a tricky issue and there does not appear to be any silver bullet approach that addresses all the concerns. Three metrics have been proposed for evaluating coreference performance over an unrestricted set of entity types: i) The **link** based MUC metric (Vilain et al., 1995), ii) The **mention** based B-CUBED metric (Bagga and Baldwin, 1998) and iii) The **entity** based CEAF (Constrained Entity Aligned F-measure) metric (Luo, 2005). Very recently BLANC (BiLateral Assessment of Noun-Phrase Coreference) measure (Recasens and Hovy, 2011) has been proposed as well. Each of the metric tries to address the shortcomings or biases of the earlier metrics. Given a set of key entities $\mathcal{K}$, and a set of response entities $\mathcal{R}$, with each entity comprising one or more mentions, each metric generates its variation of a precision and recall measure. The MUC measure if the oldest and most widely used. It focuses on the **links** (or, pairs of mentions) in the data.[18] The number of common links between entities in $\mathcal{K}$ and $\mathcal{R}$ divided by the number of links in $\mathcal{K}$ represents the recall, whereas, precision is the number of common links between entities in $\mathcal{K}$ and $\mathcal{R}$ divided by the number of links in $\mathcal{R}$. This metric prefers systems that have more mentions per entity – a system that creates a single entity of all the mentions will get a 100% recall without significant degradation in its precision. And, it ignores recall for singleton entities, or entities with only one mention. The B-CUBED metric tries to addresses MUCS's shortcomings, by focusing on the **mentions** and computes recall and precision scores for each mention. If K is the key entity containing mention M, and R is the response entity containing mention M, then recall for the mention M is computed as $\frac{|\mathrm{K}\cap\mathrm{R}|}{|\mathrm{K}|}$ and precision for the same is is computed as $\frac{|\mathrm{K}\cap\mathrm{R}|}{|\mathrm{R}|}$. Overall recall and precision are the average of the individual mention scores. CEAF aligns every response entity with at most *one* key entity by finding the best one-to-one mapping between the entities using an entity similarity metric. This is a maximum bipartite matching problem and can be solved by the Kuhn-Munkres algorithm. This is thus a **entity** based measure. Depending on the similarity, there are two variations – *entity* based CEAF – CEAF_e and a *mention* based CEAF – CEAF_e. Recall is the total similarity divided by the number of mentions in $\mathcal{K}$, and precision is the total similarity divided by the number of mentions in $\mathcal{R}$. Finally, BLANC uses a variation on the Rand index (Rand, 1971) suitable for evaluating coreference. There are a few other measures – one being the ACE value, but since this is specific to a restricted set of entities (ACE types), we did not consider it.

4.5.2 Official Evaluation Metric

In order to determine the best performing system in the shared task, we needed to associate a single number with each system. This could have been one of the metrics above, or some combination of more than one of them. The choice was not simple, and while we consulted various researchers in

[17] as is the case in this evaluation with Gold Mentions

[18] The MUC corpora did not tag single mention entities.

the field, hoping for a strong consensus, their conclusion seemed to be that each metric had its pros and cons. We settled on the MELA metric by Denis and Baldridge (2009), which takes a weighted average of three metrics: MUC, B-CUBED, and CEAF. The rationale for the combination is that each of the three metrics represents a different important dimension, the MUC measure being based on links, the B-CUBED based on mentions, and the CEAF based on entities. For a given task, a weighted average of the three might be optimal, but since we don't have an end task in mind, we decided to use the unweighted mean of the three metrics as the score on which the winning system was judged. We decided to use CEAF_e instead of CEAF_m.

4.5.3 Scoring Metrics Implementation

We used the same core scorer implementation[19] that was used for the SEMEVAL-2010 task, and which implemented all the different metrics. There were a couple of modifications done to this scorer after it was used for the SEMEVAL-2010 task.

1. Only exact matches were considered correct. Previously, for SEMEVAL-2010 non-exact matches were judged partially correct with a 0.5 score if the heads were the same and the mention extent did not exceed the gold mention.

2. The modifications suggested by Cai and Strube (2010) were incorporated in the scorer.

Since there are differences in the version used for CoNLL and the one available on the download site, and it is possible that the latter would be revised in the future, we have archived the version of the scorer on the CoNLL-2011 task webpage.[20]

5 Systems and Results

About 65 different groups demonstrated interest in the shared task by registering on the task webpage. Of these, 23 groups submitted system outputs on the test set during the evaluation week. 18 groups submitted only closed track results, 3 groups only open track results, and 2 groups submitted both closed and open track results. 2 participants in the closed track, did not write system papers, so we don't use their results in the discussion. Their results will be reported on the task webpage.

[19] http://www.lsi.upc.edu/ esapena/downloads/index.php?id=3

[20] http://conll.bbn.com/download/scorer.v4.tar.gz

The official results for the 18 systems that submitted closed track outputs are shown in Table 12, with those for the 5 systems that submitted open track results in Table 13. The official ranking score, the arithmetic mean of the F-scores of MUC, B-CUBED and CEAF_e, is shown in the rightmost column. For convenience, systems will be referred to here using the first portion of the full name, which is unique within each table.

For completeness, the tables include the raw precision and recall scores from which the F-scores were derived. The tables also include two additional scores (BLANC and CEAF_m) that did not factor into the official ranking score. Useful further analysis may be possible based on these results beyond the preliminary results presented here.

As discussed previously in the task description, we will consider three different test input conditions: i) Predicted only (Official), ii) Predicted plus gold mention *boundaries*, and iii) Predicted plus gold *mentions*

5.1 Predicted only (Official)

For the official test, beyond the raw source text, coreference systems were provided only with the predictions from automatic engines as to the other annotation layers (parses, semantic roles, word senses, and named entities).

In this evaluation it is important to note that the mention detection score cannot be considered in isolation of the coreference task as has usually been the case. This is mainly owing to the fact that there are no singleton entities in the OntoNotes data. Most systems removed singletons from the response as a post-processing step, so not only will they not get credit for the singleton entities that they correctly removed from the data, but they will be penalized for the ones that they accidentally linked with another mention. What this number does indicate is the ceiling on recall that a system would have got in absence of being penalized for making mistakes in coreference resolution. A close look at the Table 12 indicates a possible outlier in case of the *sapena* system. The recall for this system is very high, and precision way lower than any other system. Further investigations uncovered that the reason for this aberrant behavior was that fact that this system opted to *keep* singletons in the response. By design, the scorer removes singletons that might be still present in the system, but it does so *after* the mention detection accuracy is computed.

The official scores top out in the high 50's. While this is lower than the figures cited in previous coref-

System	Mention Detection			MUC			B-CUBED			CEAF$_m$			CEAF$_e$			BLANC			Official
	R	P	F	R	P	F^1	R	P	F^2	R	P	F	R	P	F^3	R	P	F	$\frac{F^1+F^2+F^3}{3}$
lee	75.07	66.81	**70.70**	61.76	57.53	59.57	68.40	68.23	68.31	56.37	56.37	**56.37**	43.41	47.75	**45.48**	70.63	76.21	73.02	**57.79**
sapena	92.39	28.19	43.20	56.32	63.16	59.55	62.75	72.08	67.09	53.51	53.51	53.51	44.75	38.38	41.32	69.50	73.07	71.10	55.99
chang	68.08	61.96	64.88	57.15	57.15	57.15	67.14	70.53	**68.79**	54.40	54.40	54.40	41.94	41.94	41.94	71.19	77.09	**73.71**	55.96
nugues	69.87	68.08	68.96	60.20	57.10	58.61	66.74	64.23	65.46	51.45	51.45	51.45	38.09	41.06	39.52	71.99	70.31	71.11	54.53
santos	67.80	63.25	65.45	59.21	54.30	56.65	68.79	62.81	65.66	49.54	49.54	49.54	35.86	40.21	37.91	73.37	66.91	69.46	53.41
song	57.81	80.41	67.26	53.73	67.79	**59.95**	60.65	66.05	63.23	46.29	46.29	46.29	43.37	30.71	35.96	69.49	59.71	61.47	53.05
stoyanov	70.84	64.98	67.78	63.61	54.04	58.43	72.58	53.27	61.44	46.08	46.08	46.08	32.00	40.82	35.88	73.21	58.93	60.88	51.92
sobha	67.82	62.09	64.83	51.08	49.88	50.48	62.63	65.43	64.00	49.48	49.48	49.48	40.65	41.82	41.23	61.40	68.35	63.88	51.90
kobdani	62.06	60.04	61.03	55.64	51.50	53.49	69.66	62.43	65.85	42.70	42.70	42.70	32.33	35.40	33.79	61.86	63.51	62.61	51.04
zhou	61.08	63.59	62.31	45.65	52.79	48.96	57.14	72.91	64.07	47.53	47.53	47.53	43.19	36.79	39.74	61.10	73.94	64.72	50.92
charton	65.90	62.77	64.30	55.09	50.05	52.45	66.26	58.44	62.10	46.82	46.82	46.82	34.33	39.05	36.54	69.94	62.23	64.80	50.36
yang	71.92	57.53	63.93	59.91	46.43	52.31	71.64	55.14	62.32	46.55	46.55	46.55	30.28	42.39	35.33	71.11	61.75	64.63	49.99
hao	64.50	64.11	64.30	57.89	51.42	54.47	67.83	55.43	61.01	45.07	45.07	45.07	30.08	35.76	32.67	72.61	62.37	65.35	49.38
xinxin	65.49	58.71	61.92	48.54	44.85	46.62	61.59	62.28	61.93	44.75	44.75	44.75	35.19	38.62	36.83	63.04	65.83	64.27	48.46
zhang	55.35	68.25	61.13	42.03	55.62	47.88	52.57	73.05	61.14	44.46	44.46	44.46	42.00	30.28	35.19	62.84	69.22	65.21	48.07
kummerfeld	69.77	56.97	62.72	46.39	39.56	42.70	63.60	57.30	60.29	45.35	45.35	45.35	35.05	42.26	38.32	58.74	61.58	59.91	47.10
zhekova	67.49	37.60	48.29	28.87	20.66	24.08	67.14	56.67	61.46	40.43	40.43	40.43	31.57	41.21	35.75	52.77	57.05	53.77	40.43
irwin	17.06	61.09	26.67	12.45	50.60	19.98	35.07	89.90	50.46	31.68	31.68	31.68	45.84	17.38	25.21	51.48	56.83	51.12	31.88

Table 12: Performance of systems in the *official, closed* track using all predicted information

System	Mention Detection			MUC			B-CUBED			CEAF$_m$			CEAF$_e$			BLANC			Official
	R	P	F	R	P	F^1	R	P	F^2	R	P	F	R	P	F^3	R	P	F	$\frac{F^1+F^2+F^3}{3}$
lee	74.31	67.87	**70.94**	62.83	59.34	**61.03**	68.85	69.01	**68.93**	56.70	56.70	**56.70**	43.29	46.80	**44.98**	71.90	76.55	73.96	**58.31**
cai	67.15	67.64	67.40	56.73	58.90	57.80	64.60	71.03	67.66	53.37	53.37	53.37	42.71	40.68	41.67	69.77	73.96	71.62	55.71
uryupina	70.60	66.31	68.39	59.70	55.70	57.63	66.29	64.12	65.18	51.42	51.42	51.42	38.34	42.17	40.16	69.23	68.54	68.88	54.32
klenner	64.41	60.28	62.28	49.04	50.71	49.86	61.70	68.61	64.97	50.03	50.03	50.03	41.28	39.70	40.48	66.05	73.90	69.05	51.77
irwin	24.60	62.27	35.27	18.56	51.01	27.21	38.97	85.57	53.55	33.86	33.86	33.86	43.33	19.36	26.76	51.62	52.91	51.76	35.84

Table 13: Performance of systems in the *official, open* track using all predicted information

System	Mention Detection			MUC			B-CUBED			CEAF$_m$			CEAF$_e$			BLANC			Official
	R	P	F	R	P	F^1	R	P	F^2	R	P	F	R	P	F^3	R	P	F	$\frac{F^1+F^2+F^3}{3}$
lee	79.52	71.25	**75.16**	65.87	62.05	**63.90**	69.52	70.55	**70.03**	59.26	59.26	**59.26**	46.29	50.48	**48.30**	72.00	78.55	**74.77**	**60.74**
nugues	74.18	70.74	72.42	64.33	60.05	62.12	68.26	65.17	66.68	53.84	53.84	53.84	39.86	44.23	41.93	72.53	71.04	71.75	56.91
chang	63.37	73.18	67.92	55.00	65.50	59.79	62.16	76.65	68.65	54.95	54.95	54.95	46.77	37.17	41.42	70.97	79.30	74.29	56.62
santos	65.82	69.90	67.80	57.76	61.39	59.52	64.49	70.27	67.26	51.87	51.87	51.87	41.42	38.16	39.72	72.72	71.97	72.34	55.50
kobdani	67.11	65.09	66.08	62.63	56.80	59.57	73.20	62.22	67.27	44.49	44.49	44.49	32.87	37.25	34.92	64.07	64.13	64.10	53.92
stoyanov	76.90	64.73	70.29	69.81	55.01	61.54	77.07	52.54	62.48	48.08	48.08	48.08	30.97	44.84	36.64	76.57	60.33	62.96	53.55
zhang	59.62	71.19	64.89	46.06	58.75	51.64	53.89	73.41	62.16	46.62	46.62	46.62	43.49	32.11	36.95	64.11	70.47	66.54	50.25
song	58.43	77.64	66.68	46.66	68.40	55.48	54.40	70.19	61.29	43.62	43.62	43.62	43.77	25.88	32.53	66.29	58.76	60.22	49.77
zhekova	69.19	57.27	62.67	33.48	37.15	35.22	55.47	68.23	61.20	41.31	41.31	41.31	38.29	34.65	36.38	53.45	63.33	54.79	44.27

Table 14: Performance of systems in the supplementary *closed* track using predicted information plus *gold boundaries*

System	Mention Detection			MUC			B-CUBED			CEAF$_m$			CEAF$_e$			BLANC			Official
	R	P	F	R	P	F^1	R	P	F^2	R	P	F	R	P	F^3	R	P	F	$\frac{F^1+F^2+F^3}{3}$
lee	78.71	72.33	75.39	66.93	63.91	65.39	70.09	71.49	70.78	59.78	59.78	59.78	46.34	49.62	47.92	73.38	79.00	75.83	61.36

Table 15: Performance of systems in the supplementary *open* track using predicted information plus *gold boundaries*

System	Mention Detection			MUC			B-CUBED			CEAF_m			CEAF_e			BLANC			Official
	R	P	F	R	P	F^1	R	P	F^2	R	P	F	R	P	F^3	R	P	F	$\frac{F^1+F^2+F^3}{3}$
chang	100	100	100	80.46	84.75	82.55	72.84	74.57	73.70	69.71	69.71	69.71	70.45	60.75	65.24	78.01	76.57	77.26	73.83

Table 16: Performance of systems in the *supplementary, closed* track using predicted information plus *gold mentions*

System	Mention Detection			MUC			B-CUBED			CEAF_m			CEAF_e			BLANC			Official
	R	P	F	R	P	F^1	R	P	F^2	R	P	F	R	P	F^3	R	P	F	$\frac{F^1+F^2+F^3}{3}$
lee	83.37	100	90.93	74.79	89.68	81.56	67.46	86.88	75.95	70.73	70.73	70.73	77.75	51.05	61.64	76.65	85.85	80.35	73.05

Table 17: Performance of systems in the *supplementary, open* track using predicted information plus *gold mentions*

System	Mention Detection			MUC			B-CUBED			CEAF_m			CEAF_e			BLANC			Official
	R	P	F	R	P	F^1	R	P	F^2	R	P	F	R	P	F^3	R	P	F	$\frac{F^1+F^2+F^3}{3}$
lee	76.79	68.34	**72.32**	63.29	58.96	61.05	68.84	68.72	68.78	57.28	57.28	**57.28**	44.19	48.75	**46.36**	70.93	76.58	73.36	**58.73**
sapena	95.27	29.07	44.55	56.99	63.91	60.25	62.89	72.31	67.27	53.90	53.90	53.90	45.22	38.70	41.71	69.71	73.32	71.32	56.41
chang	69.88	63.61	66.60	58.48	58.48	58.48	67.42	70.91	**69.12**	55.21	55.21	55.21	42.66	42.66	42.66	71.42	77.36	**73.96**	56.75
nugues	72.96	71.08	72.01	62.68	59.46	61.03	67.24	64.89	66.04	52.82	52.82	52.82	39.25	42.50	40.81	72.57	70.86	71.68	55.96
santos	70.39	65.67	67.95	61.28	56.20	58.63	69.25	63.16	66.07	50.47	50.47	50.47	36.51	41.15	38.69	73.92	67.32	69.93	54.46
song	59.24	82.39	68.92	54.92	69.29	61.27	60.89	66.27	63.46	46.97	46.97	46.97	44.49	31.15	36.65	69.73	59.87	61.61	53.79
stoyanov	74.43	68.28	71.22	67.18	57.08	**61.72**	74.06	53.45	62.09	47.40	47.40	47.40	32.78	42.52	37.02	74.10	59.34	61.31	53.61
sobha	71.06	65.06	67.93	53.91	52.64	53.27	63.17	66.14	64.62	50.80	50.80	50.80	41.77	43.03	42.39	61.91	69.15	64.49	53.43
kobdani	65.98	63.83	64.89	59.22	54.81	56.93	70.49	63.12	66.60	44.17	44.14	44.15	33.19	36.50	34.77	62.52	64.25	63.32	52.77
zhou	64.11	66.74	65.40	48.00	55.51	51.48	57.18	73.71	64.40	48.40	48.40	48.40	44.18	37.35	40.48	61.54	74.86	65.30	52.12
charton	71.01	67.64	69.28	59.24	53.82	56.40	67.10	59.02	62.80	48.91	48.91	48.91	35.96	41.39	38.48	70.65	62.71	65.34	52.56
yang	73.73	58.97	65.53	61.23	47.45	53.47	71.88	55.13	62.40	47.05	47.05	47.05	30.54	43.16	35.77	71.39	61.92	64.83	50.55
hao	66.79	66.38	66.59	59.55	52.89	56.02	68.27	55.46	61.20	45.95	45.95	45.95	30.76	36.81	33.51	73.22	62.73	65.78	50.24
xinxin	69.05	61.91	65.28	50.99	47.11	48.97	61.59	62.70	62.14	45.64	45.64	45.64	35.86	39.57	37.62	63.42	66.29	64.68	49.58
zhang	57.41	70.78	63.40	43.48	57.53	49.53	52.44	73.60	61.24	44.97	44.97	44.97	42.71	30.44	35.55	63.12	69.63	65.53	48.77
kummerfeld	71.05	58.01	63.87	47.42	40.44	43.65	63.73	57.39	60.39	45.76	45.76	45.76	35.30	42.72	38.66	58.89	61.77	60.07	47.57
zhekova	72.65	40.48	51.99	31.73	22.70	26.46	66.92	56.68	61.37	41.04	41.04	41.04	31.93	42.17	36.34	53.09	57.86	54.22	41.39
irwin	17.58	62.96	27.49	12.69	51.59	20.37	34.88	89.98	50.27	31.71	31.71	31.71	46.13	17.33	25.20	51.51	56.93	51.14	31.95

Table 18: *Head word based* performance of systems in the *official, closed* track using all predicted information

System	Mention Detection			MUC			B-CUBED			CEAF_m			CEAF_e			BLANC			Official
	R	P	F	R	P	F^1	R	P	F^2	R	P	F	R	P	F^3	R	P	F	$\frac{F^1+F^2+F^3}{3}$
lee	76.01	69.43	**72.57**	64.40	60.83	**62.57**	69.34	69.57	**69.45**	57.68	57.68	**57.68**	44.15	47.85	**45.92**	72.23	76.94	**74.32**	**59.31**
cai	69.32	69.82	69.57	58.39	60.63	59.49	64.88	71.53	68.04	54.36	54.36	54.36	43.74	41.58	42.64	70.13	74.39	72.01	56.72
uryupina	72.10	67.72	69.84	60.74	56.68	58.64	66.43	64.25	65.32	52.00	52.00	52.00	38.87	42.85	40.76	69.43	68.73	69.07	54.91
klenner	71.73	67.14	69.36	55.17	57.04	56.09	62.67	70.69	66.44	53.25	53.25	53.25	44.27	42.39	43.31	67.45	75.92	70.68	55.28
irwin	25.24	63.87	36.18	18.90	51.94	27.71	38.79	85.64	53.40	33.89	33.89	33.89	43.59	19.31	26.76	51.66	52.98	51.80	35.96

Table 19: *Head word based* performance of systems in the *official, open* track using all predicted information

erence evaluations, that is as expected, given that the task here includes predicting the underlying mentions and mention boundaries, the insistence on exact match, and given that the relatively easier appositive coreference cases are not included in this measure. The top-performing system (*lee*) had a score of 57.79 which is about 1.8 points higher than that of the second (*sapena*) and third (*chang*) ranking systems, which scored 55.99 and 55.96 respectively. Another 1.5 points separates them from the fourth best score of 54.53 (*nugues*). Thus the performance differences between the better-scoring systems were not large, with only about three points separating the top four systems.

This becomes even clearer if we merge in the results of systems that participated only in the open track but that made relatively limited use of outside resources.[21] Comparing that way, the *cai* system scores in the same ball park as the second rank systems (*sapena* and *chang*). The *uryupina* system similarly scores very close to *nugues*'s 54.53

Given that our choice of the official metric was somewhat arbitrary, if is also useful to look at the individual metrics, including the mention-based CEAF_m and BLANC metrics that were not part of the official metric. The *lee* system which scored the best using the official metric does slightly worse than *song* on the MUC metric, and also does slightly worse than *chang* on the B-CUBED and BLANC metrics. However, it does much better than every other group on the entity-based CEAF_e, and this is the primary reason for its 1.8 point advantage in the official score. If the CEAF_e measure does indicate the accuracy of entities in the response, this suggests that the *lee* system is doing better on getting coherent entities than any other system. This could be partly due to the fact that that system is primarily a precision-based system that would tend to create purer entities. The CEAF_e measure also seems to penalize other systems more harshly than do the other measures.

We cannot compare these results to the ones obtained in the SEMEVAL-2010 coreference task using a small portion of OntoNotes data because it was only using nominal entities, and had heuristically added singleton mentions to the OntoNotes data[22]

[21]The *cai* system specifically mentions that, and the only resource that the *uryupina* system used outside of the closed track setting was the Stanford named entity tagger.

[22]The documentation that comes with the SEMEVAL data package from LDC (LDC2011T01) states: "Only nominal mentions and identical (IDENT) types were taken from the OntoNotes coreference annotation, thus excluding coreference relations with verbs and appositives. Since OntoNotes is only annotated with multi-mention entities, singleton referential elements were identified heuristically: all NPs and possessive determiners were annotated as singletons excluding those functioning as appositives or as pre-modifiers but for NPs in the possessive case. In coordinated NPs, single constituents as well as the entire NPs were considered to be mentions. There is no reliable heuristic to automatically detect English expletive pronouns, thus they were (although inaccurately) also annotated as singletons."

5.2 Predicted plus gold mention *boundaries*

We also explored performance when the systems were provided with the gold mention boundaries, that is, with the exact spans (expressed in terms of token offsets) for all of the NP constituents in the human-annotated parse trees for the test data. Systems could use this additional data to ensure that the output mention spans in their entity chains would not clash with those in the answer set. Since this was a secondary evaluation, it was an *optional* element, and not all participants ran their systems on this task variation. The results for those systems that did participate in this optional task are shown in Tables 14 (closed track) and 15 (open track).

Most of the better scoring systems did supply these results. While all systems did slightly better here in terms of raw scores, the performance was not much different from the official task, indicating that mention boundary errors resulting from problems in parsing do not contribute significantly to the final output.[23]

One side benefit of performing this supplemental evaluation was that it revealed a subtle bug in the automatic scoring routine that we were using that could double-count duplicate correct mentions in a given entity chain. These can occur, for example, if the system considers a unit-production NP-PRP combination as two mentions that identify the exact same token in the text, and reports them as separate mentions. Most systems had a filter in their processing that selected only one of these duplicate mentions, but the *kobdani* system considered both as potential mentions, and its developers tuned their algorithm using that flawed version of the scorer.

When we fixed the scorer and re-evaluated all of the systems, the *kobdani* system was the only one whose score was affected significantly, dropping by about 8 points, which lowered that system's rank from second to ninth. It is not clear how much of this was owing to the fact that the system's param-

[23]It would be interesting to measure the overlap between the entity clusters for these two cases, to see whether there was any substantial difference in the mention chains, besides the expected differences in boundaries for individual mentions.

eters had been tuned using the scorer with the bug, which double-credited duplicate mentions. To find out for sure, one would have to re-tune the system using the modified scorer.

One difficulty with this supplementary evaluation using gold mention boundaries is that those boundaries alone provide only very partial information. For the roughly 10% of mentions that the automatic parser did not correctly identify, while the systems knew the correct boundaries, they had no hierarchical parser or semantic role label information, and they also had to further approximate the already heuristic head word identification. This incomplete data complicated the systems' task and also complicates interpretation of the results.

5.3 Predicted plus gold *mentions*

The final supplementary condition that we explored was if the systems were supplied with the manually-annotated spans for exactly those mentions that did participate in the gold standard coreference chains. This supplies significantly more information than the previous case, where exact spans were supplied for all NPs, since the gold mentions list here will also include verb headwords that are linked to event NPs, but will not include singleton mentions, which do not end up as part of any chain. The latter constraint makes this test seem somewhat artificial, since it directly reveals part of what the systems are designed to determine, but it still has some value in quantifying the impact that mention detection has on the overall task and what the results are if the mention detection is perfect.

Since this was a logical extension of the task and since the data was available to the participants for the development set, a few of the sites did run experiments of this type. Therefore we decided to provide the gold *mentions* data to a few sites who had reported these scores, so that we could compute the performance on the test set. The results of these experiments are shown in Tables 16 and 17. The results show that performance does go up significantly, indicating that it is markedly easier for the systems to generate better entities given gold *mentions*. Although, ideally, one would expect a perfect mention detection score, it is the case that one of the two systems – *lee* – did not get a 100% Recall. This could possibly be owing to unlinked singletons that were removed in post-processing.

The *lee* system developers also ran a further experiment where both gold mentions for the elements of the coreference chains and also gold *annotations* for all the other layers were available to the system. Surprisingly, the improvement in coreference performance from having gold annotation of the other layers was almost negligible. This suggests that either: i) the automatic models are predicting those layers well enough that switching to gold doesn't make much difference; ii) information from the other layers does not provide much leverage for coreference resolution; or iii) current coreference models are not capable of utilizing the information from these other layers effectively. Given the performance numbers on the individual layers cited earlier, (i) seems unlikely, and we hope that further research in how best to leverage these layers will result in models that can benefit from them more definitively.

5.4 Head word based scoring

In order to check how stringent the *official*, exact match scoring is, we also performed a relaxed scoring. Unlike ACE and MUC, the OntoNotes data does not have manually annotated minimum spans that a mention must contain to be considered correct. However, OntoNotes does have manual syntactic analysis in the form of the Treebank. Therefore, we decided to approximate the minimum spans by using the head words of the mentions using the gold standard syntax tree. If the response mention contained the head word and did not exceed the true mention boundary, then it was considered correct – both from the point of view of mention detection, and coreference resolution. The scores using this relaxed strategy for the *open* and *closed* track submissions using predicted data are shown in Tables 18 and 19. It can be observed that the relaxed, head word based, scoring does not improve performance very much. The only exception was the *klenner* system whose performance increased from 51.77 to 55.28. Overall, the ranking remained quite stable, though it did change for some adjacent systems which had very close *exact match* scores.

5.5 Genre variation

In order to check how the systems did on various genres, we scored their performance per genre as well. Tables 20 and 21 summarize genre based performance for the *closed* and *open* track participants respectively. System performance does not seem to vary as much across the different genres as is normally the case with language processing tasks, which could suggest that coreference is relatively genre insensitive, or it is possible that scores are two low for the difference to be apparent. Comparisons are difficult, however, because the spoken gen-

		MD F	MUC F	BCUB F	C_m F	C_e F	BLANC F	O F
lee	GENRE							
	BC	72.2	60.0	66.2	53.9	43.7	71.7	56.7
	BN	72.0	59.0	68.7	57.6	48.7	68.8	58.8
	MZ	70.1	58.0	72.2	61.6	50.9	75.0	60.4
	NW	65.4	54.3	69.4	56.5	45.5	70.4	56.4
	TC	75.9	66.8	69.5	59.3	41.3	81.6	59.2
	WB	73.0	63.9	65.7	54.2	42.7	73.4	57.5
sapena								
	BC	48.7	58.8	64.6	50.8	39.4	70.4	54.3
	BN	47.1	60.0	69.1	57.4	45.0	74.3	58.0
	MZ	35.3	59.2	72.3	60.4	48.2	75.0	59.9
	NW	35.2	57.9	69.7	55.3	41.9	73.8	56.5
	TC	60.4	64.3	63.3	48.3	35.1	68.8	54.2
	WB	46.3	60.1	62.5	49.1	37.4	67.4	53.3
chang								
	BC	65.5	56.4	67.1	51.5	39.8	71.6	54.4
	BN	66.6	57.4	69.1	56.0	45.6	70.5	57.4
	MZ	61.6	52.7	71.3	57.6	46.4	72.9	56.8
	NW	61.0	53.3	69.1	54.1	42.1	71.9	54.8
	TC	72.2	68.5	71.4	59.6	37.7	81.7	59.2
	WB	66.4	59.7	66.7	52.7	39.4	74.7	55.3
nugues								
	BC	71.4	59.2	62.4	48.2	37.2	68.4	52.9
	BN	70.0	58.5	67.4	54.5	43.1	73.1	56.3
	MZ	65.4	53.6	68.6	54.2	42.2	70.1	54.8
	NW	61.8	51.9	67.0	51.3	39.2	69.4	52.7
	TC	77.2	69.2	63.9	53.0	37.9	72.2	57.0
	WB	72.9	64.2	63.4	51.1	38.5	74.3	55.4
santos								
	BC	66.6	57.2	64.8	48.5	37.2	68.6	53.0
	BN	66.9	57.3	66.9	52.3	41.0	71.8	55.1
	MZ	62.7	51.0	65.9	48.9	37.8	64.5	51.6
	NW	58.4	49.5	66.2	48.1	37.4	66.9	51.0
	TC	74.2	66.9	65.9	52.5	35.5	72.5	56.1
	WB	70.4	63.2	63.4	49.5	38.2	70.3	55.0
song								
	BC	68.9	61.4	61.0	44.1	34.3	59.5	52.2
	BN	66.2	58.4	64.8	49.0	38.2	65.2	53.8
	MZ	63.7	53.4	65.5	49.9	39.0	63.4	52.6
	NW	62.4	53.6	64.3	48.0	37.2	62.7	51.7
	TC	76.9	74.4	62.0	43.3	33.2	58.1	56.5
	WB	70.0	63.0	60.1	43.3	31.8	60.8	51.6
stoyanov								
	BC	69.5	59.1	57.6	43.5	34.0	58.7	50.2
	BN	69.2	59.1	65.4	50.4	40.0	65.5	54.8
	MZ	66.7	55.1	65.5	51.0	39.9	63.7	53.5
	NW	61.8	52.0	63.3	46.2	36.1	62.0	50.5
	TC	72.6	66.6	57.6	42.3	31.0	57.6	51.7
	WB	71.5	63.9	58.3	44.8	33.1	61.1	51.8
sobha								
	BC	68.3	51.7	61.4	47.8	40.4	62.9	51.2
	BN	66.5	51.9	66.5	53.7	45.5	66.3	54.6
	MZ	68.8	54.9	70.3	58.9	49.3	69.8	58.1
	NW	55.1	43.1	65.8	48.6	39.0	64.9	49.3
	TC	71.5	55.1	57.5	44.2	36.7	60.5	49.7
	WB	70.5	55.7	59.2	46.6	39.8	62.6	51.6
kobdani								
	BC	63.2	56.3	65.8	40.6	32.4	61.9	51.5
	BN	63.5	55.7	68.5	46.9	37.5	64.6	53.9
	MZ	57.5	52.2	69.8	45.7	36.4	61.7	52.8
	NW	52.2	41.7	64.4	43.2	33.7	62.6	46.6
	TC	67.7	60.2	65.3	36.6	28.5	57.6	51.3
	WB	68.7	62.8	62.4	42.5	32.9	64.0	52.7
zhou	GENRE							
	BC	64.1	49.5	62.1	45.3	38.8	61.8	50.1
	BN	60.8	45.9	64.4	49.5	41.2	66.8	50.5
	MZ	58.8	44.4	66.9	50.1	41.8	64.6	51.0
	NW	57.7	44.8	65.7	48.7	40.3	63.1	50.2
	TC	69.2	58.1	60.8	43.1	35.7	62.6	51.5
	WB	67.4	55.4	62.8	47.9	39.2	69.1	52.5
charton								
	BC	65.8	53.1	59.1	44.6	35.2	64.4	49.1
	BN	65.5	52.0	64.0	50.0	39.6	65.9	51.9
	MZ	61.7	46.3	64.6	49.7	39.9	64.1	50.3
	NW	57.6	44.6	64.5	48.2	37.7	67.0	48.9
	TC	73.1	66.8	56.2	42.8	29.9	58.1	51.0
	WB	67.6	57.6	59.3	45.1	33.3	66.6	50.0
yang								
	BC	65.7	53.8	62.3	46.8	35.0	67.5	50.3
	BN	66.0	53.1	63.8	49.1	40.0	63.1	52.3
	MZ	58.8	43.9	59.7	42.6	32.8	55.5	45.5
	NW	57.2	44.7	62.9	45.3	35.0	62.7	47.6
	TC	74.2	66.8	66.3	55.3	36.0	76.1	56.4
	WB	67.6	57.6	57.0	42.6	32.1	60.1	48.9
hao								
	BC	68.9	58.7	58.9	44.8	31.7	64.9	49.8
	BN	62.0	51.1	63.0	46.2	35.5	64.1	49.9
	MZ	60.3	46.7	61.5	46.3	34.3	61.9	47.5
	NW	57.2	47.7	63.3	45.5	32.9	66.0	48.0
	TC	67.9	60.4	58.8	44.7	30.3	68.3	49.8
	WB	71.4	61.8	55.7	42.6	30.0	64.4	49.2
xinxin								
	BC	64.8	47.8	60.2	43.9	35.5	65.1	47.9
	BN	61.5	44.7	63.2	47.0	38.9	65.8	48.9
	MZ	54.6	35.5	64.5	45.7	37.7	61.0	45.9
	NW	54.3	39.5	64.0	45.0	37.5	61.1	47.0
	TC	74.2	62.0	57.9	45.4	33.4	66.5	51.1
	WB	66.9	52.6	58.5	42.2	35.9	63.4	49.0
zhang								
	BC	65.8	50.6	61.1	45.3	35.5	67.3	49.1
	BN	56.3	43.9	61.0	45.8	35.8	66.8	46.9
	MZ	57.1	35.1	62.2	44.4	36.1	59.4	44.5
	NW	49.9	37.8	61.8	43.2	35.2	59.8	44.9
	TC	75.4	65.9	60.2	46.0	32.1	67.1	52.7
	WB	69.2	55.4	57.4	42.5	34.6	64.7	49.1
kummerfield								
	BC	66.4	41.5	55.6	41.7	36.2	57.9	44.4
	BN	68.3	48.2	63.4	51.7	44.7	61.6	52.1
	MZ	58.0	39.9	65.8	51.0	43.4	64.1	49.7
	NW	55.2	41.3	64.7	46.8	37.0	63.5	47.6
	TC	61.8	34.5	51.5	34.7	30.0	54.1	38.7
	WB	68.2	48.1	56.0	44.4	38.6	59.6	47.6
zhekova								
	BC	50.5	23.8	60.6	39.4	35.1	53.4	39.8
	BN	51.2	26.0	62.4	42.5	37.5	54.3	42.0
	MZ	44.0	22.6	63.4	43.3	37.3	56.0	41.1
	NW	39.7	19.4	62.8	41.0	35.8	53.7	39.3
	TC	59.4	31.6	58.2	37.7	33.6	54.1	41.1
	WB	54.1	27.8	58.7	38.5	34.7	53.0	40.4
irwin								
	BC	23.5	16.1	46.0	29.4	23.6	49.8	28.6
	BN	24.9	20.0	49.7	34.2	27.1	52.9	32.3
	MZ	23.2	17.9	55.9	36.2	28.5	53.0	34.1
	NW	27.5	21.6	56.4	33.9	27.3	52.6	35.1
	TC	28.0	19.3	38.2	24.5	18.7	49.0	25.4
	WB	33.6	24.8	47.6	29.7	23.0	50.2	31.8

Table 20: Detailed look at the performance per *genre* for the *official, closed* track using automatic performance. MD represents MENTION DETECTION; BCUB represents B-CUBED; C_m represents $CEAF_m$; C_e represents $CEAF_e$ and O represents the OFFICIAL score.

res were treated here with perfect speech recognition accuracy and perfect speaker turn information. Under more realistic application conditions, the spread in performance between genres might be greater.

		MD F	MUC F	BCUB F	C_m F	C_e F	BLANC F	O F
lee	GENRE							
	BC	72.7	61.7	67.0	54.5	43.6	72.7	57.4
	BN	72.0	60.6	69.4	57.9	48.1	70.3	59.3
	MZ	69.9	58.4	72.1	61.2	50.1	75.2	60.2
	NW	65.3	55.8	70.0	56.7	44.9	71.7	56.9
	TC	76.6	68.4	70.4	59.6	40.8	82.1	59.9
	WB	73.8	65.5	66.2	54.5	42.1	74.2	57.9
cai								
	BC	69.7	59.1	66.0	50.5	39.9	69.2	55.0
	BN	68.6	57.6	67.8	55.4	45.5	68.2	56.9
	MZ	64.0	51.1	69.5	55.9	45.6	71.2	55.4
	NW	60.3	49.9	67.8	52.7	41.2	69.1	53.0
	TC	75.6	70.5	72.2	59.6	38.0	80.3	60.2
	WB	71.7	63.9	65.0	51.8	39.8	72.8	56.2
uryupina								
	BC	70.2	58.3	62.7	48.7	38.0	68.7	53.0
	BN	69.0	57.6	66.8	53.6	43.1	69.2	55.8
	MZ	65.7	52.4	68.3	54.3	43.6	68.8	54.8
	NW	62.6	52.1	68.3	53.2	41.2	71.3	53.9
	TC	75.7	67.1	61.0	50.7	34.6	67.1	54.2
	WB	72.0	61.7	60.9	48.8	38.3	67.6	53.6
klenner								
	BC	63.2	50.3	63.4	48.2	38.9	66.8	50.8
	BN	63.1	48.6	65.0	51.0	42.6	66.0	52.1
	MZ	59.1	43.7	67.1	52.9	45.3	65.0	52.0
	NW	55.3	41.3	65.0	48.0	39.6	64.5	48.7
	TC	73.9	64.9	67.9	56.4	39.0	78.0	57.3
	WB	66.8	58.1	64.0	50.1	39.6	72.7	53.9
irwin								
	BC	36.6	27.6	50.9	32.0	25.5	50.2	34.7
	BN	30.8	24.6	51.9	36.4	28.6	54.8	35.0
	MZ	26.1	20.0	57.3	37.6	29.4	54.3	35.6
	NW	32.3	24.7	58.4	34.7	27.9	51.1	37.0
	TC	46.4	34.3	44.6	29.4	21.9	51.7	33.6
	WB	41.7	32.9	50.5	32.9	25.1	53.2	36.2

Table 21: Detailed look at the performance per *genre* for the *official, open* track using predicted information. MD represents MENTION DETECTION; BCUB represents B-CUBED; C_m represents $CEAF_m$; C_e represents $CEAF_e$ and O represents the OFFICIAL score.

6 Approaches

Tables 22 and 23 summarize the approaches of the participating systems along with some of the important dimensions.

Most of the systems broke the problem into two phases, first identifying the potential mentions in the text and then linking the mentions to form coreference chains. Most participants also used rule-based approaches for mention detection, though two did use trained models. While trained morels seem able to better balance precision and recall, and thus to achieve a higher F-score on the mention task itself, their recall tends to be quite a bit lower than that achievable by rule-based systems designed to favor recall. This impacts coreference scores because the full coreference system has no way to recover if the mention detection stage misses a potentially anaphoric mention.

Only one of the participating systems *cai* attempted to do joint mention detection and coreference resolution. While it did not happen to be among the top-performing systems, the difference in performance could be due to the richer features used by other systems rather than to the use of a joint model.

Most systems represented the markable mentions internally in terms of the parse tree NP constituent span, but some systems used shared attribute models, where the attributes of the merged entity are determined collectively by heuristically merging the attribute types and values of the different constituent mentions.

Various types of trained models were used for predicting coreference. It is interesting to note that some of the systems, including the best-performing one, used a completely rule-based approach even for this component.

Most participants appear not to have focused much on eventive coreference, those coreference chains that build off verbs in the data. This usually meant that mentions that should have linked to the eventive verb were instead linked in with some other entity. Participants may have chosen not to focus on events because they pose unique challenges while making up only a small portion of the data. Roughly 91% of mentions in the data are NPs and pronouns.

In the systems that used trained models, many systems used the approach described in Soon et al. (2001) for selecting the positive and negative training examples, while others used some of the alternative approaches that have been introduced in the research literature more recently. Many of the trained systems also were able to improve their performance by using feature selection, though things varied some depending on the example selection strategy and the classifier used. Almost half of the trained systems used the feature selection strategy from Soon et al. (2001) and found it beneficial. It is not clear whether the other systems did not explore this path, or whether it just did not prove as useful in their case.

7 Conclusions

In this paper we described the anaphoric coreference information and other layers of annotation in the

	Task	Syntax	Learning Framework	Markable Identification	Markable	Verb	Feature Selection	# Features	Training
lee	C+O	P	Rule-based	Rules to exclude Copular construction, Appositives, Pleonastic *it*, etc.	Feature dependent with shared attributes	×	×		—
sapena	C	P	Decision Tree + Relaxation Labeling	NP (maximal span) + PRP + NE + Capitalized noun heuristic	Full phrase	×	×		Train + Dev
chang	C	P	Learning Based Java	NP, NE, PRP, PRP$	Full phrase	×	×		Train + Dev
cai	O	P	Compute hyperedge weights on 30% of training data	NP, PRP, PRP$, Base phrase chunks, Pleonastic *it* filter	Full phrase	×	×		—
nugues	C	D	Logistic Regression (LIBLINEAR)	NP, PRP$ and sequence of NNP(S) in post processing using ALIAS and STRINGMATCH	Head word	×	Forward + Backward starting from Soon feature set	24	Train + Dev
uryupina	O	P	Decision Tree. Different classifiers for Pronominal and non-Pronominal mentions	NP, NE, PRP, PRP$, and rules to exclude some specific cases	Full phrase	×	Multi-Objective Optimization on three splits. NSGA-II	46	Train + Dev
santos	C	P	ETL (Entropy guided Transformational Learning) committee and Random Forest (WEKA)	All NP and all pronouns and PER, ORG, GPE in NP	Full phrase	×	Inherent to the classifiers		Train + Dev
song	C	P	MaxEnt (OpenNLP)	Mention detection classifier	Full phrase	×	Same feature set, but per classifier	40	Train
stoyanov	C	P	Averaged perceptron	NE and possessives in addition to ACE based system	Full phrase	×	×	76	—
sobha	C	P	CRF for non-pronominal and salience factor for pronominal resolution	Machine learned pleonastic *it*, plus NP, PRP, PRP$ and NE	Minimal (Chunk/NE) and Maximum span	×	×		Train
klenner	O	D	Rule-based. Salience measure using dependencies generated from training data	NP, NE, PRP, PRP$	Shared attributed/transitivity by using a virtual prototype	×	×		—
kobdani	C	P	Decision Tree	NP (no mention of PRP$)	Start word, End word and Head of NP	×	Information gain ratio		Train
zhou	C	P	SVM tree kernel using BC portion of the data	Rule-based; Five rules: PRP$, PRP, NE, smallest NP subsuming NE and DET+NP	Full phrase	×	×	17	Train + Dev
charton	C	P	Multi-layer perceptron	Rules based on POS, NE and filter out pleonastic *it* using rule-based filter	Full phrase	×	×	22	Train
yang	C	P	MaxEnt (MALLET)	NP, PRP, PRP$, pre-modifiers and verbs	Full phrase	√	×	40	Train + Dev
hao	C	P	MaxEnt	NP, PRP, PRP$, VBD	full phrase	√	×		Train + Dev
xinxin	C	P	ILP/Information gain	NP, PRP, PRP$	Full phrase	×	Information gain ratio	65	—
zhang	C	P	SVM	IOB classification	Full phrase	×	×		—
kummerfield	C	P	Unsupervised generative model	NP, PRP, PRP$ with maximal span	Full phrase	×	×		—
zhekova	C	P	TIMBL memory based learner	NP, Proper nouns, PRP, PRP$, plus verb with predicate lemma	Head word	√	×		Train + Dev
irwin	C+O	P	Classification-based ranker	NP, PRP, PRP$	Shared attributes	×	×		—

Table 22: Participating system profiles – Part I. In the Task column, C/O represents whether the system participated in the *closed*, *open* or both tracks. In the Syntax column, a P represents that the systems used a phrase structure grammar representation of syntax, whereas a D represents that they used a dependency representation.

	Positive Training Examples	Negative Training Examples	Decoding	Parse Configuration
lee	—	—	Multi-pass Sieves	
sapena	All mention pairs and longer of nested mentions with common head kept	Mention pairs with less than threshold (5) number of different attribute values are considered (22% out of 99% original are discarded)	Iterative	1-best
chang	Closest antecedent	All preceding mentions in a union of of *gold* and *predicted* mentions. Mentions where the first is pronoun and other not are not considered	Best link and All links strategy; with and without constraints – Best link without constraints was selected for the official run	
cai	Weights are trained on part of the training data		Recursive 2-way Spectral clustering (Agarwal, 2005)	
nugues	Closest Antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	Closest-first clustering for pronouns and Best-first clustering for non-pronouns	1-best
uryupina	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	mention pair model without ranking as in Soon 2001	
santos	Extended version of Soon (2001) where in addition to their strategy, positive and negative examples from mentions in the sentence of the closest preceding antecedent are considered		Limited number of preceding mentions 60 for automatic and 40 given gold boundaries; Aggressive-merge clustering (Mccarthy and Lenhert, 1995)	
song	Pre-cluster pair models separate for each pair NP-NP, NP-PRP and PRP-PRP		Pre-clusters, with singleton pronoun pre-clusters, and use closest-first clustering. Different link models based on the type of linking mentions – NP-PRP, PRP-PRP and NP-NP	
stoyanov	Smart Pair Generation (SmartPG) where the type of antecedent is determined by the type of anaphor using a set of rules		Single-link clustering by computing transitive closure between pairwise positives.	
sobha	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	Pronominal: all preceding NPs in the sentence and preceding 4 sentences	
klenner	—	—	Incremental entity creation	
kobdani	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	Best-first clustering. Threshold of 100 words used for long documents	1-best
zhou	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	—	
charton	From the end of the document, until an antecedent is found, or 10 mentions	Negative examples in between anaphor and closest antecedent	MLP with score of 0.5 used for linking and 10 mentions	
yang	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	Maximum 23 sentences to the left; Constrained clustering	
hao	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	Beam search (Luo, 2004)	Packed forest
xinxin	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	Best-first clustering followed by ILP optimization	
zhang	Closest antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	Window of 100 markables	
kummerfield	—	—	Pre- and post- resolution filters	Given + Berkeley parser parses; parses without NMLs improved performance slightly; re-trained Berkeley parser
zhekova	Examples in the past three sentences		From last possible mention in document	
irwin	Cluster query with NULL cluster for discourse new mentions		Cluster-ranking approach (rahman, 2009)	

Table 23: Participating system profiles – Part II. This focuses on the way positive and negative examples were generated and the decoding strategy used.

OntoNotes corpus, and presented the results from an evaluation on learning such unrestricted entities and events in text. The following represent our conclusions on reviewing the results:

- Perhaps the most surprising finding was that the best-performing system (*lee*) was completely rule-based, rather than trained. This suggests that their rule-based approach was able to do a more effective job of combining the multiple sources of evidence than the trained systems. The features for coreference prediction are certainly more complex than for many other language processing tasks, which makes it more challenging to generate effective feature combinations. The rule-based approach used by the best-performing system seemed to benefit from a heuristic that captured the most confident links before considering less confident ones, and also made use of the information in the guidelines in a slightly more refined manner than other systems. They also included appositives and copular constructions in their calculations. Although OntoNotes does not count those as instances of IDENT coreference, using that information may have helped their system discover additional useful links.

- It is interesting to note that the developers of the *lee* system also did the experiment of running their system using gold standard information on the individual layers, rather than automatic model predictions. The somewhat surprising result was that using perfect information for the other layers did not end up improving coreference performance much, if at all. It is not clear whether this means that: i) Automatic predictors for the individual layers are accurate enough already; ii) Information captured by those supplementary layers actually does not provide much leverage for resolving coreference; or iii) researchers have yet have found an effective way of capturing and utilizing the extra information provided by these layers.

- It does seem that collecting information about an entity by merging information across the various attributes of the mentions that comprise it can be useful, though not all systems that attempted this achieved a benefit.

- System performance did not seem to vary as much across the different genres as is normally the case with language processing tasks, which could suggest that coreference is relatively genre insensitive, or it is possible that scores are two low for the difference to be apparent. Comparisons are difficult, however, because the spoken genres were treated here with perfect speech recognition accuracy and perfect speaker turn information. Under more realistic application conditions, the spread in performance between genres might be greater.

- It is noteworthy that systems did not seem to attempt the kind of joint inference that could make use of the full potential of various layers available in OntoNotes, but this could well have been owing to the limited time available for the shared task.

- We had expected to see more attention paid to event coreference, which is a novel feature in this data, but again, given the time constraints and given that events represent only a small portion of the total, it is not surprising that most systems chose not to focus on it.

- Scoring coreference seems to remain a significant challenge. There does not seem to be an objective way to establish one metric in preference to another in the absence of a specific application. On the other hand, the system rankings do not seem terribly sensitive to the particular metric chosen. It is interesting that both versions of the CEAF metric – which tries to capture the goodness of the entities in the output – seem much lower than the other metric, though it is not clear whether that means that our systems are doing a poor job of creating coherent entities or whether that metric is just especially harsh.

Finally, it is interesting to note that the problem of coreference does not seem to be following the same kind of learning curve that we are used to with other problems of this sort. While performance has improved somewhat, it is not clear how far we will be able to go given the strategies at hand, or whether new techniques will be needed to capture additional information from the texts or from world knowledge. We hope that this corpus and task will provide a useful resource for continued experimentation to help resolve this issue.

Acknowledgments

We gratefully acknowledge the support of the Defense Advanced Research Projects Agency

(DARPA/IPTO) under the GALE program, DARPA/CMO Contract No. HR0011-06-C-0022. We would like to thank all the participants. Without their hard work, patience and perseverance this evaluation would not have been a success. We would also like to thank the Linguistic Data Consortium for making the OntoNotes 4.0 corpus freely and timely available to the participants. Emili Sapena, who graciously allowed the use of his scorer implementation, and made available enhancements and immediately fixed issues that were uncovered during the evaluation. Finally, we offer our special thanks to Lluís Màrquez and Joakim Nivre for their wonderful support and guidance without which this task would not have been successful.

References

Olga Babko-Malaya, Ann Bies, Ann Taylor, Szuting Yi, Martha Palmer, Mitch Marcus, Seth Kulick, and Libin Shen. 2006. Issues in synchronizing the English treebank and propbank. In *Workshop on Frontiers in Linguistically Annotated Corpora 2006*, July.

Amit Bagga and Breck Baldwin. 1998. Algorithms for Scoring Coreference Chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 33–40, Sydney, Australia, July.

Jie Cai and Michael Strube. 2010. Evaluation metrics for end-to-end coreference resolution systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 28–36.

Eugene Charniak and Mark Johnson. 2001. Edit detection and parsing for transcribed speech. In *Proceedings of the Second Meeting of North American Chapter of the Association of Computational Linguistics*, June.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and maxent discriminative reranking. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, MI, June.

Nancy Chinchor and Beth Sundheim. 2003. Message understanding conference (MUC) 6. In *LDC2003T13*.

Nancy Chinchor. 2001. Message understanding conference (MUC) 7. In *LDC2001T02*.

Aron Culotta, Michael Wick, Robert Hall, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *HLT/NAACL*, pages 81–88.

Pascal Denis and Jason Baldridge. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *Proceedings of HLT/NAACL*.

Pascal Denis and Jason Baldridge. 2009. Global joint models for coreference resolution and named entity classification. *Procesamiento del Lenguaje Natural*, (42):87–96.

Charles Fillmore, Christopher Johnson, and Miriam R. L. Petruck. 2003. Background to framenet. *International Journal of Lexicography*, 16(3).

G. G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassell, and R. Weischedel. 2000. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *Proceedings of LREC*.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393, Los Angeles, California, June.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 1–18, Boulder, Colorado, June.

Sanda M. Harabagiu, Razvan C. Bunescu, and Steven J. Maiorano. 2001. Text and knowledge mining for coreference resolution. In *NAACL*.

L. Hirschman and N. Chinchor. 1997. Coreference task definition (v3.0, 13 jul 97). In *Proceedings of the Seventh Message Understanding Conference*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of HLT/NAACL*, pages 57–60, New York City, USA, June. Association for Computational Linguistics.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2000. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21 – 40.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical*

Methods in Natural Language Processing, pages 25–32, Vancouver, British Columbia, Canada, October.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19(2):313–330, June.

Andrew McCallum and Ben Wellner. 2004. Conditional models of identity uncertainty with application to noun coreference. In *Advances in Neural Information Processing Systems (NIPS)*.

Joseph McCarthy and Wendy Lehnert. 1995. Using decision trees for coreference resolution. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence*, pages 1050–1055.

Thomas S. Morton. 2000. Coreference for nlp applications. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, October.

Vincent Ng. 2007. Shallow semantics for coreference resolution. In *Proceedings of the IJCAI*.

Vincent Ng. 2010. Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411, Uppsala, Sweden, July.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Martha Palmer, Hoa Trang Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically.

R. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of LREC*.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*.

Massimo Poesio. 2004. The mate/gnome scheme for anaphoric annotation, revisited. In *Proceedings of SIGDIAL*.

Simone Paolo Ponzetto and Massimo Poesio. 2009. State-of-the-art nlp approaches to coreference resolution: Theory and practical recipes. In *Tutorial Abstracts of ACL-IJCNLP 2009*, page 6, Suntec, Singapore, August.

Simone Paolo Ponzetto and Michael Strube. 2005. Semantic role labeling for coreference resolution. In *Companion Volume of the Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics*, pages 143–146, Trento, Italy, April.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the HLT/NAACL*, pages 192–199, New York City, N.Y., June.

Sameer Pradhan, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James Martin, and Dan Jurafsky. 2005. Support vector learning for semantic argument classification. *Machine Learning Journal*, 60(1):11–39.

Sameer Pradhan, Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2007a. OntoNotes: A Unified Relational Semantic Representation. *International Journal of Semantic Computing*, 1(4):405–419.

Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007b. Unrestricted Coreference: Indentifying Entities and Events in OntoNotes. In *in Proceedings of the IEEE International Conference on Semantic Computing (ICSC)*, September 17-19.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 968–977, Singapore, August. Association for Computational Linguistics.

W. M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336).

Marta Recasens and Eduard Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*.

Marta Recasens, Lluís Màrquez, Emili Sapena, M. Antònia Martí, Mariona Taulé, Véronique Hoste, Massimo Poesio, and Yannick Versley. 2010. Semeval-2010 task 1: Coreference resolution in multiple languages. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 1–8, Uppsala, Sweden, July.

W. Soon, H. Ng, and D. Lim. 2001. A machine learning approach to coreference resolution of noun phrase. *Computational Linguistics*, 27(4):521–544.

Veselin Stoyanov, Nathan Gilbert, Claire Cardie, and Ellen Riloff. 2009. Conundrums in noun phrase coreference resolution: Making sense of the state-of-the-art. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 656–664, Suntec, Singapore, August. Association for Computational Linguistics.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings*

of the Twelfth Conference on Computational Natural Language Learning, pages 159–177, Manchester, England, August.

Yannick Versley. 2007. Antecedent selection techniques for high-recall coreference resolution. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

M. Vilain, J. Burger, J. Aberdeen, D. Connolly, and L. Hirschman. 1995. A model theoretic coreference scoring scheme. In *Proceedings of the Sixth Message Undersatnding Conference (MUC-6)*, pages 45–52.

Ralph Weischedel and Ada Brunstein. 2005. BBN pronoun coreference and entity type corpus LDC catalog no.: LDC2005T33. BBN Technologies.

Ralph Weischedel, Eduard Hovy, Martha Palmer, Mitch Marcus, Robert Belvin, Sameer Pradhan, Lance Ramshaw, and Nianwen Xue. 2011. OntoNotes: A Large Training Corpus for Enhanced Processing. In Joseph Olive, Caitlin Christianson, and John McCary, editors, *Handbook of Natural Language Processing and Machine Translation*. Springer.