CoNLL-2012 Shared Task:
Modeling Multilingual Unrestricted Coreference in OntoNotes

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

Alessandro Moschitti
University of Trento,
38123 Povo (TN)
Italy
moschitti@disi.unitn.it

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

Olga Uryupina
University of Trento,
38123 Povo (TN)
Italy
uryupina@gmail.com

Yuchen Zhang
Brandeis University,
Waltham, MA 02453
USA
yuchenz@brandeis.edu

Abstract

The CoNLL-2012 shared task involved predicting coreference in English, Chinese, and Arabic, using the final version, v5.0, of the OntoNotes corpus. It was a follow-on to the English-only task organized in 2011. Until the creation of the OntoNotes corpus, resources in this sub-field of language processing were limited to noun phrase coreference, often on a restricted set of entities, such as the 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, and covers multiple languages. OntoNotes also provides additional layers of integrated annotation, capturing additional shallow semantic structure. This paper describes the OntoNotes annotation (coreference and other layers) and then describes the parameters of the shared task including the format, pre-processing information, evaluation criteria, and presents and discusses the results achieved by the participating systems. The task of coreference has had a complex evaluation history. Potentially many evaluation conditions, have, in the past, made it difficult to judge the improvement in new algorithms over previously reported results. Having a standard test set and standard evaluation parameters, all based on a resource that provides multiple integrated annotation layers (syntactic parses, semantic roles, word senses, named entities and coreference) and in multiple languages could support joint modeling and help ground and 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.

Early work on corpus-based coreference resolution dates back to the mid-90s by McCarthy and Lenhert (1995) where they experimented with decision trees and hand-written rules. Corpora to support supervised learning of this task date back to the Message Understanding Conferences (MUC) (Hirschman and Chinchor, 1997; Chinchor, 2001; Chinchor and Sundheim, 2003). The de facto standard datasets for current coreference studies are the MUC and the ACE[1] (Doddington et al., 2004) corpora. These corpora were tagged with coreferring entities in the form of noun phrases in the text. The MUC corpora cover all noun phrases in text but are relatively small in size. The ACE corpora, on the other hand, cover much more data, but the annotation is restricted to a small subset of entities.

Automatic identification of coreferring entities and events in text has been an uphill battle for several decades, partly because it is a problem that requires world knowledge to solve and word knowledge is hard to define, and partly owing to the lack of substantial annotated data. Aside from the fact that resolving coreference in text is simply a very hard problem, there have been other hindrances that further contributed to the slow progress in this area:

(i) *Smaller sized corpora* such as MUC which covered coreference across all noun phrases. Corpora such as ACE which are larger in size, *but cover a smaller set of entities*; and

(ii) *low consistency in existing corpora* annotated with coreference — in terms of inter-annotator agreement (ITA) (Hirschman et al., 1998) — owing to attempts at covering multiple coreference phenomena that are not equally annotatable with high agreement which likely lessened the reliability of statistical evidence in the form of lexical coverage and semantic relatedness that could be derived from the data and

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

Proceedings of the Joint Conference on EMNLP and CoNLL: Shared Task, pages 1–40,
Jeju Island, Korea, July 13, 2012. ©2012 Association for Computational Linguistics

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 to take language understanding applications such as question answering or distillation to the next level, we need more consistent annotation for larger amounts of broad coverage data to train better automatic models for entity and event detection.

(iii) *Complex evaluation* with multiple evaluation metrics and multiple evaluation scenarios, complicated with varying training and test partitions, led to situations where many researchers report results with only one or a few of the available metrics and under a subset of evaluation scenarios. This has made it hard to gauge the improvement in algorithms over the years (Stoyanov et al., 2009), or to determine which particular areas require further attention. 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 relatively easy (Culotta et al., 2007).

(iv) *the knowledge bottleneck* which has been a well-accepted ceiling that has kept the progress in this task at bay.

These issues suggest that the following steps might take the community in the right direction towards improving the state of the art in coreference resolution:

(i) Create a *large corpus* with *high inter-annotator agreement* possibly by restricting the coreference annotating to phenomena that can be annotated with high consistency, and *covering an unrestricted set of entities and events*; and

(ii) Create a *standard evaluation scenario* with an official evaluation setup, and possibly several ablation settings to capture the range of performance. This can then be used as a standard benchmark by the research community.

(iii) Continue to *improve learning algorithms* that better incorporate world knowledge and jointly incorporate information from other layers of syntactic and semantic annotation to improve the state of the art.

One of the many goals of the OntoNotes project[2] (Hovy et al., 2006; Weischedel et al., 2011) was to explore whether it could 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 subset of entity types. The coreference layer in OntoNotes constitutes just one part of a multi-layered, integrated annotation of shallow semantic structures in text with high inter-annotator agreement. This addresses the first issue.

In the language processing community, the field of speech recognition probably has the longest history of shared evaluations held primary by NIST[3] (Pallett, 2002). In the past decade machine translation has been a topic of shared evaluations also by NIST[4]. There are many syntactic and semantic processing tasks that are not quite amenable to such continued evaluation efforts. The CoNLL shared tasks over the past 15 years have filled that gap, helping establish benchmarks and advance the state of the art in various sub-fields within NLP. The importance of shared tasks is now in full display in the domain of clinical NLP (Chapman et al., 2011) and recently a coreference task was organized as part of the i2b2 workshop (Uzuner et al., 2012). The computational learning community is also witnessing a shift towards joint inference based evaluations, with the two previous CoNLL tasks (Surdeanu et al., 2008; Hajič et al., 2009) devoted to joint learning of syntactic and semantic dependencies. A SemEval-2010 coreference task (Recasens et al., 2010) was the first attempt to address the second issue. It included six different Indo-European languages — Catalan, Dutch, English, German, Italian, and Spanish. Among other corpora, a small subset (∼120K) of English portion of OntoNotes was used for this purpose. However, the lack of a strong participation prevented the organizers from reaching any firm conclusions. The CoNLL-2011 shared task was another attempt to address the second issue. It was well received, but the shared task was only *limited to the English portion of OntoNotes.* In addition, the coreference portion of OntoNotes did not have a concrete baseline prior to the 2011 evaluation, thereby making it challenging for participants to gauge the performance of their algorithms in the absence of established state of the art on this flavor of annotation. The closest comparison was to the results reported by Pradhan et al. (2007b) on the newswire portion of OntoNotes. Since the corpus also covers two other languages from completely different language families, Chinese and Arabic, it provided a great opportunity to have a *follow-on task in 2012 covering all*

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

[3] http://www.itl.nist.gov/iad/mig/publications/ASRhistory/index.html

[4] http://www.itl.nist.gov/iad/mig/tests/mt/

three languages. As we will see later, peculiarities of each of these languages had to be considered in creating the evaluation framework.

The first systematic learning-based study in coreference resolution was conducted on the MUC corpora, using a decision tree learner, by Soon et al. (2001). Significant improvements have been made in the field of language processing in general, and improved learning techniques have pushed 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). Researchers have continued to find novel ways of exploiting ontologies such as WordNet. Various knowledge sources from shallow semantics to encyclopedic knowledge have been exploited (Ponzetto and Strube, 2005; Ponzetto and Strube, 2006; Versley, 2007; Ng, 2007). 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). More recently researchers have used graph based algorithms (Cai et al., 2011a) rather than pair-wise classifications. For a detailed survey of the progress in this field, we refer the reader to a recent article (Ng, 2010) and a tutorial (Ponzetto and Poesio, 2009) dedicated to this subject. 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. Further research to reduce the knowledge gap is essential to take coreference resolution techniques to the next level.

The rest of the paper is organized as follows: Section 2 presents an overview of the OntoNotes corpus. Section 3 describes the range of phenomena annotated in OntoNotes, and language-specific issues. Section 4 describes the shared task data and the evaluation parameters, with Section 4.4.2 examining the performance of the state-of-the-art tools on all/most intermediate layers of annotation. Section 5 describes the participants in the task. Section 6 briefly compares the approaches taken by various participating systems. Section 7 presents the system results with some analysis. Section 8 compares the performance of the systems on the a subset of the Engish test set that corresponds with the test set used for the CoNLL-2011 evaluation. Section 9 draws some conclusions.

2 The OntoNotes Corpus

The OntoNotes project has created a large-scale corpus of accurate and integrated annotation of multiple levels of the shallow semantic structure in text. The English and Chinese language portion comprises roughly one million words per language of newswire, magazine articles, broadcast news, broadcast conversations, web data and conversational speech data. The English subcorpus also contains an additional 200K words of the English translation of the New Testament as Pivot Text. The Arabic portion is smaller, comprising 300K words of newswire articles. The hope is that this rich, integrated annotation covering many layers will allow for richer, cross-layer models and enable significantly better automatic semantic analysis. In addition to coreference, this data is also tagged with syntactic trees, propositions for most verb and some noun instances, partial verb and noun word senses, and 18 named entity types. Manual annotation of a large corpus with multiple layers of syntax and semantic information is a costly endeavor. Over the years in the development of this corpus, there were various priorities that came into play, and therefore not all the data in the corpus could be annotated with all the different layers of annotation. However, such multi-layer annotations, with complex, cross-layer dependencies, demands a robust, efficient, scalable storage mechanism 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 facilitates the extraction of cross-layer features in integrated predictive models that will make use of these annotations.

OntoNotes comprises the following layers of annotation:

- **Syntax —** A layer of syntactic annotation for English, Chinese and Arabic based on a revised guidelines for the Penn Treebank (Marcus et al., 1993; Babko-Malaya et al., 2006), the Chinese Treebank (Xue et al., 2005) and the Arabic Treebank (Maamouri and Bies, 2004).
- **Propositions —** The proposition structure of verbs based on revised guidelines for the English PropBank (Palmer et al., 2005; Babko-Malaya et al., 2006), the Chinese PropBank (Xue and Palmer, 2009) and the Arabic PropBank (Palmer et al., 2008; Zaghouani et al., 2010).
- **Word Sense —** Coarse-grained word senses are tagged for the most frequent polysemous verbs and nouns, in order to maximize token

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. In case of English and Arabic languages, the sense-inventories (and frame files) are defined separately for each part of speech that is realized by the lemma in the text. For Chinese, however the sense inventories (and frame files) are defined per lemma — independent of the part of speech realized in the text. For the English portion of OntoNotes, each individual sense has been connected to multiple WordNet senses. This provides users direct access to the WordNet semantic structure. There is also a mapping from the OntoNotes word senses to PropBank frames and to VerbNet (Kipper et al., 2000) and FrameNet (Fillmore et al., 2003). Unfortunately, owing to lack of comparable resources as comprehensive as WordNet in Chinese or Arabic, neither language has any inter-resource mappings available.

- **Named Entities —** The corpus was tagged with a set of 18 well-defined proper named entity types that have been tested extensively 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). It considers all pronouns (PRP, PRP$), noun phrases (NP) and heads of verb phrases (VP) as potential mentions. Unlike English, Chinese and Arabic have dropped subjects and objects which were also considered during coreference annotation[5]. 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 in the OntoNotes corpus. Two different types of coreference are distinguished: Identity (IDENT), and Appositive (APPOS). Identity coreference (IDENT) 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. Appositives (APPOS) are treated separately because they function as attributions, as described further below. 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 guidelines are fairly language independent. We will look at some salient aspects of the coreference annotation in OntoNotes. For more details, and examples, we refer the reader to the release documentation. We will primarily use English examples to describe various aspects of the annotation and use Chinese and Arabic examples especially to illustrate phenomena not observed in English, or that have some language specific peculiarities.

3.1 Noun Phrases

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 parse trees in the syntactic layer of OntoNotes annotation, though the annotators can also add additional mentions when appropriate. In the following two examples (and later ones), the phrases 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.

Noun phrases (NPs) in Chinese can be complex noun phrases or bare nouns (nouns that lack a determiner such as "the" or "this"). Complex noun phrases contain structures modifying the head noun, as in the following examples:

(3) (他担任 总统 任内 最后 一 次 的 (亚 太 经 济 合作 会议 (高峰会))).
((His last APEC (summit meeting)) as the President)

(4) (越南 统一 后 (第一 位 前往 当地 访问 的 (美国 总统)))
((The first (U.S. president)) who went to visit Vietnam after its unification)

In these examples, the smallest phrase in parentheses is the bare noun. The longer phrase in parantheses includes modifying structures. All the expressions in the parantheses, however, share the same head noun, i.e., "高峰会 (summit meeting)", and "美国总统 (U.S. president)" respectively. Nested noun phrases, or nested NPs, are contained within

[5] As we will see later these are not used during the task.

longer noun phrases. In the above example, "summit meeting" and "U.S. president" are nested NPs. Wherever NPs are nested, the largest logical span is used in coreference.

3.2 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 the single-word verb head is marked for convenience. This includes morphologically related nominalizations as in (5) and noun phrases that refer to the same event, even if they are lexically distinct from the verb as in (6). In the following two examples, only the chains related to the *growth* event are shown in bold. The Arabic translation of the same example identifies mentions using parantheses.

(5) The European economy **grew** rapidly over the past years, **this growth** helped raising ....

لقد (نما) الإقتصاد الأوروبي بسرعة خلال السنوات
الماضية، (هذا النمو) ساهم في رفع ...

(6) 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.

3.3 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 in the coreference annotation.)

(7) 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.

In Chinese, all the following pronouns — 你，我，他, 她, 它， 你们，我们，他们，它们，我, 您, 咱们 (*you, me, he, she*, and so on), and demonstrative pronouns — 这个，那个， 这些, 那些 (*this, that, these, those*) in singular, plural or possessive forms are linked to anything they refer to.

Pronouns from classical Chinese such as 其中 (*among which*), 其 (*he/she/it*), 之 (*he/she/it*) are also linked with other mentions to which they refer.

In Arabic, the following pronouns are coreferenced – nominative personal pronouns (subject) and demonstrative pronouns which are detached. Subject pronouns are often null in Arabic; overt subject pronouns are rare, but do occur.

هما / هم / هن / نحن / انتما / انتم / انتن
(*We, you, they*)

انا / انت / هو / هي
(*I, you, she, he*)

Object pronouns are attached to the verb (direct objects) or preposition (indirect objects)

ي / ك / ـه / ها
(*Me, you, him, her*)

نا / كُم / كُن / كُما / هُم / هُن
(*Us, you, them*)

and, possessive (adjectival) pronouns are identical to object pronouns, but are attached to nouns.

ي / ك / ـه / ها
(*My, your, his, her*)

نا / كُم / كُن / كُما / هُم / هُن
(*Our, your, their*)

Pronouns such as 你，您，你们，大家，各位 can be considered generic. In this case, they are not linked to other generic mentions in the same discourse. For example,

(8) 请 ***大家** 带好自己的随身物品。***大家** 请下车。
Please take your belongings with ***you**. Please get off the train, ***everyone**.

In Chinese, if the subject or object can be recovered from the context, or if it is of little interest for the reader/listener to know, it can be omitted. In the Chinese Treebank, a small *pro* in inserted in positions where the subject or object is omitted. A *pro* can be replaced by overt NPs if they refer to the same entity or event, and the *pro* and its overt NP antecedent do not have to be in the same sentence. Exactly what *pro* stands for is determined by the linguistic context in which it appears.

(9) 吉林省主管经贸工作的副省长全哲洙说：“(***pro***) 欢迎国际社会同(我们) 一道，共同推进图门江开发事业， 促进区域经济发展，造福东北亚人民。
Quan Zhezhu, Vice Governor of Jinlin Province who is in charge of economics and trade, said: "(***pro***) Welcome international societies to join (us) in the development of Tumen Jiang, so as to promote regional economic development and benefit people in Northeast Asia.

Sometimes, *pro*s cannot be recovered in the text—i.e., an overt NP cannot be identified as their antecedent in the same text — and therefore they are not linked. For instance, the *pro* in existential sentences usually cannot be recovered or linked in the annotation, as in the following example:

(10) **(*pro*)** 有二十三项高新技术项目进区开发。
There are 23 high-tech projects under development in the zone.

Also, if *pro* does not refer to a specific entity or event, it is considered generic *pro* and not linked as in (11).

(11) 肯德基 、 麦当劳 等 速食店 全 大陆 都 推出 了 **(*pro*)** 买 套餐赠送 布质 或 棉质 圣诞 老人 玩具 的 促销.
In Mainland China, fast food restaurants such as Kentucky Fried Chicken and McDonald's have launched their promotional packages by providing free cotton Santa toys for each combo **(*pro*)** purchased.

Finally, *pro*s in idiomatic expressions are not linked. Similar to Chinese, Arabic null subjects and objects are also eligible for coreference and treated similarly. In the Arabic Treebank, these are marked with just an "*". There exists few of these instances in English — marked (yet differently) with a *PRO* in the treebank and which are connected in PropBank annotation but not in coreference.

3.4 Generic mentions

Generic nominal mentions can be linked with referring pronouns and other definite mentions, but not with other generic nominal mentions.

This would allow linking of the bolded mentions in (12) and (13), but not in (14).

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

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

(14) 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 (12) and (13), are always considered generic. In example (15) below, there are three 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.

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

(16) 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.

3.5 Pre-modifiers

Proper pre-modifiers can be coreferenced, but proper nouns that are in a morphologically adjectival form are treated as adjectives, and are 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* in "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.

(17) **FBI** spokesman

(18) ***U.S.** spokesman

In Chinese adjectival and nominal forms of GPEs are not morphologically distinct, and in such cases the annotator decides whether it is an adjectival usage. Usually if something is tagged as NORP then it is not considered as a mention.

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

(19) 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.

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

3.6 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 proposition annotation.

(21) **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 are not linked. Since *Called* is copular, neither IDENT nor APPOS coreference is marked in the following case.

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

Some examples of copular verbs in Chinese are 是 (*to be*) and 为 (*to be, to serve as*). In addition, other verbs (particularly so-called *light verbs*) that trigger an attributive reading on the following NP: 成为 (*become*), (当)选为 (*is elected*), 称为 (*is called*), (好)像 (*looks like*), 叫做 (*is called*), etc.

(23) **(上海)**是***(中国最大的城市)**。**(上海)**发展得很快。
(Shanghai) is ***(the largest city in China)**. **(Shanghai)** develops fast.

In the above example, the two mentions of 上海 (*Shanghai*) co-refer with each other, but the entity does not co-refer with 中国最大的城 市 (*the largest city in China*).

3.7 Small clauses

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

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

Note that the mention *Fred*, however, can be connected to other mentions of *Fred* in the text.

3.8 Temporal expressions

Temporal expressions such as the following are linked:

(25) 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*.

(26) 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 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.

3.9 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 capturing the attributed property even though there is no explicit copula.

(27) **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:

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

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

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
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.

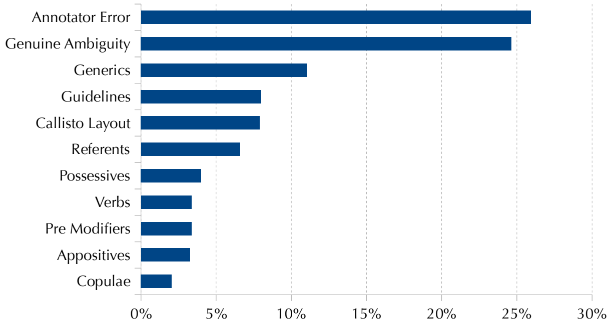

Figure 2: The distribution of disagreements across the various types in Table 1 for a sample of 15K disagreements in the English portion of the corpus.

(30) **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:

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

(32) The sheriff, his friend

(33) 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.

(34) A dangerous bacteria, bacillium, is found

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.

(35) **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*):

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

Similar rules apply for Chinese and Arabic. Unlike English, where most appositives have a punctuation marker, in Chinese that is not necessarily the frequent case. In the following example we can see an appositive construction without any punctuations between the head and the attribute.

(37) 上图左起：(无锡市市长)$_{\text{X[attribute]}}$ (王宏民)$_{\text{X[head]}}$， (副市长)$_{\text{Y[attribute]}}$ (洪锦、张怀西)$_{\text{Y[head]}}$， …

Language	Genre	A1-A2	A1-ADJ	A2-ADJ
English	Newswire [NW]	80.9	85.2	88.3
	Broadcast News [BN]	78.6	83.5	89.4
	Broadcast Conversation [BC]	86.7	91.6	93.7
	Magazine [MZ]	78.4	83.2	88.8
	Weblogs and Newsgroups [WB]	85.9	92.2	91.2
	Telephone Conversation [TC]	81.3	94.1	84.7
	Pivot Text [PT] (New Testament)	89.4	96.0	92.0
Chinese	Newswire [NW]	73.6	84.8	75.1
	Broadcast News [BN]	80.5	86.4	91.6
	Broadcast Conversation [BC]	84.1	90.7	91.2
	Magazine [MZ]	74.9	81.2	80.0
	Weblogs and Newsgroups [WB]	87.6	92.3	93.5
	Telephone Conversation [TC]	65.6	86.6	77.1

Table 1: Inter Annotator (A1 and A2) and Adjudicator (ADJ) agreement for the Coreference Layer in OntoNotes measured in terms of the MUC score.

Figure above from left : **Wuxi Mayor**$_{\text{X[attribute]}}$ **Wang Hongmin**$_{\text{X[head]}}$, **Deputy Mayors**$_{\text{Y[attribute]}}$ **Hong Jin, Zhang Huaixi**$_{\text{Y[head]}}$, ...

3.10 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.
- GPEs are linked to references to their governments, even when the references are nested NPs, or the modifier and head of a single NP.
- In extremely rare cases, metonymic mentions can be co-referenced. This is done only when the two mentions clearly and without a doubt refer to the same entity. For example:

  (38) In a statement released this afternoon, **10 Downing Street** called the bombings in Casablanca "a strike against all peace-loving people."

  (39) In a statement, **Britain** called the Casablanca bombings "a strike against all peace-loving people."

  In this case, it is obvious that "10 Downing Street" and "Britain" are being used interchangeably in the text. Again, if there is any ambiguity, however, these terms are not coreferenced with each other.
- In Arabic, verbal inflections are not considered pronominal and are not coreferenced. The portion marked with an * in the example below is an inflection and not a pronoun, and so should not be marked.

  (40) صرح (**ت***) الناطقة ب آسم وزارة الخارجية
  السويسرية دانييلا ستوفل : آ إن (**ها**) ليست
  في برن و لا في جنيف

  The Swiss foreign ministry's spokeswoman announced the (**she**) is neither in Burne nor in Geneva Pronouns in quoted speech are also marked.

3.11 Annotator Agreement and Analysis

Table 1 shows the inter-annotator and annotator-adjudicator agreement on all the genres and languages of OntoNotes. A 15K disagreements in various parts of the English data was analyzed, and grouped 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. Interestingly, this mirrors the annotator disagreement analysis on the MUC corpus provided by Hirschman et al. (1998).

4 CoNLL-2012 Coreference Task

The CoNLL-2012 shared task was held across all three languages — English, Chinese and Arabic — of the OntoNotes v5.0 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 coreference decisions had to be made using automatically predicted information on other structural and semantic layers including the parses, semantic roles, word senses, and named entities. Given various factors, such as the lack of resources and state-of-the-art tools, and time constraints, we could not provide some layers of information for the Chinese and Arabic portion of the data.

The three languages are from quite different language families. The morphology of these languages is quite different. Arabic has a complex morphology, English has limited morphology, whereas Chinese has very little morphology. English word segmentation amounts to rule-based tokenization, and is close to perfect. In the case of Chinese and Arabic, although the tokenization/segmentation is not as good as English, the accuracies are in the high 90s. Syntactically, there are many dropped subjects and objects in Arabic and Chinese, whereas English is not a pro-drop language. Another difference is the amount of resources available for each language. English has probably the most resources at its disposal, whereas Chinese and Arabic lack significantly

— Arabic more so than Chinese. Given this fact, plus the fact that the CoNLL format cannot handle multiple segmentations, and that it would complicate scoring since we are using exact token boundaries (as discussed later in Section 4.5), we decided to allow the use of gold, treebank segmentation for all languages. In the case of Chinese, the words themselves are lemmas, so no additional information needs to be provided. For Arabic, by default written text is unvocalised, so we decided to also provide correct, gold standard lemmas, along with the correct vocalized version of the tokens. Table 2 lists which layers were available and quality of the provided layers (when provided.)

Layer	English	Chinese	Arabic
Segmentation	●	●	●
Lemma	√	—	●
Parse	√	√	√[6]
Proposition	√	√	×
Predicate Frame	√	×	×
Word Sense	√	√	√
Name Entities	√	×	×
Speaker	●	●	—

Table 2: Summary of predicted layers provided for each language. A "●" indicates gold annotation, a "√" indicates predicted, a "×" indicates an absence of the predicted layer, and a "—" indicates that the layer is not applicable to the language.

As is customary for CoNLL tasks, there were two *primary* 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. Within each *closed* and *open* track, we had an optional *supplementary* track which allowed us to run some ablation studies over a few different input conditions. This allowed us to evaluate the systems given: i) Gold mention boundaries (GB), ii) Gold mentions (GM), and iii) Gold parses (GS). We will refer to the main task – where no mention boundaries are provided – as NB.

4.1 Primary Evaluation

The primary evaluation comprises the *closed* and *open* tracks where predicted information is provided on all layers of the test set other than coreference. As mentioned earlier, we provide gold lemma and vocalization information for Arabic, and we use gold standard treebank segmentation for all three languages.

4.1.1 Closed Track

In the *closed* track, systems were limited to the provided data. For the training and test data, in addition to the underlying text, *predicted* versions of all the supplementary layers of annotation were provided using off-the-shelf tools (parsers, semantic role labelers, named entity taggers, etc.) retrained on the training portion of the OntoNotes data — 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 to train 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 if they wished to use a lattice of predictions.

More so than previous CoNLL shared tasks, coreference predictions depend on world knowledge, and many state-of-the-art systems use information from external resources such as WordNet, which provides a layer of information that could help a system recognize semantic connections between the various lexicalized mentions in the text. Therefore, in the case of English, similar to the previous year's task, we allowed the use of WordNet in the closed track. Since word senses in OntoNotes are predominantly[7] coarse-grained groupings of WordNet senses, systems could also map from the predicted or gold-standard word senses 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 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. Unfortunately neither Arabic, nor Chinese have comparable resources available that we could allow participants to use. Chinese, in particular, does not have number or gender inflections for nouns, but (Baran and Xue, 2011) look at a way to infer such information.

4.1.2 Open Track

In addition to resources available in the *closed* track, in the *open* track, systems were allowed to use

[6] The predicted part of speech for Arabic are a mapped down version of the richer gold version present in the treebank

[7] 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

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

Procedure: Generate_Partitions(OntoNotes) **returns** Train, Dev, Test

```
Train $\leftarrow \emptyset$
Dev $\leftarrow \emptyset$
Test $\leftarrow \emptyset$
for all Source $\in$ OntoNotes do
  if Source = Wall Street Journal then
    Train $\leftarrow$ Train $\cup$ Sections 02 – 21
    Dev $\leftarrow$ Dev $\cup$ Sections 00, 01, 22, 24
    Test $\leftarrow$ Test $\cup$ Section 23
  else
    if Number of files in Source $\geq 10$ then
      Train $\leftarrow$ Train $\cup$ File IDs ending in 1 – 8
      Dev $\leftarrow$ Dev $\cup$ File IDs ending in 0
      Test $\leftarrow$ Test $\cup$ File IDs ending in 9
    else
      Dev $\leftarrow$ Dev $\cup$ File IDs ending in 0
      Test $\leftarrow$ Test $\cup$ File ID ending in the highest number
      Train $\leftarrow$ Train $\cup$ Remaining File IDs for the Source
    end if
  end if
end for
return Train, Dev, Test
```

external resources such as Wikipedia, gazetteers etc. The purpose of this track is mainly to get an idea of the performance ceiling on the task at the cost of not being able to perform a fair 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 — so they can participate in the task with minimal, or no modification to their existing system.

4.2 Supplementary Evaluation

In addition to the option of selecting between the primary *closed* or the *open* tracks, the participants also had an option to run their systems in the following ablation settings.

Gold Mention Boundaries (GB) In this case, we provided all possible correct mention boundaries in the test data. This essentially entails all NPs, and PRPs in the data extracted from the gold parse trees, as well as the mentions that do not align with any parse constituent, for example, non-existent constituents in the predicted parse owing to errors, some named entities, etc.

Gold Mentions (GM) In this dataset, we provided *only* and *all* the correct mentions for the test sets, thereby reducing the task to one of pure mention clustering, and eliminating the task of mention detection and anaphoricity determination[8]. These also include potential spans that do not align with any constituent in the predicted parse tree.

Gold Parses (GS) In this case, for each language, we replaced the predicted parses in the *closed* track data with manual, gold parses.

4.3 Train, Development and Test Splits

For various reasons, not all the documents in OntoNotes have been annotated with all the different layers of annotation, with full coverage.[9] There is a core portion, however, which is roughly 1.6M English words, 950K Chinese words, and 300K Arabic words which has been annotated with all the layers. This is the portion that we used for the shared task.

We used the same algorithm as in CoNLL-2011 to

[8] Mention detection interacts with anaphoricity determination since the corpus does not contain any singleton mentions.

[9] As mentioned earlier, large scale manual annotation of various layers of syntax and semantics is an expensive endeavor. Adding to this, the fact that word sense annotation is most efficiently done one lemma at a time, ideally all instances of the same across the entire corpus, or as large a portion as possible, full coverage across all lemma instances is hard to achieve given the long tail of low frequency lemmas with a Zipfian distribution. Similar issue affects PropBank annotation, but furthermore, currently it only covers mostly verb predicates, and a few eventive noun predicates.

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

[11] These numbers are for the part of OntoNotes v5.0 that have all layers of annotation including coreferenced.

Corpora	Language	Words				Documents			
		Total	Train	Dev	Test	Total	Train	Dev	Test
MUC-6	English	25K	12K	13K		60	30	30	
MUC-7	English	40K	19K	21K		67	30	37	
ACE[10](2000-2004)	English	960K	745K	215K		-	-	-	
	Chinese	615K	455K	150K		-	-	-	
	Arabic	500K	350K	150K		-	-	-	
OntoNotes[11]	English	1.6M	1.3M	160K	170K	2,384 (3493)	1,940 (2,802)	222 (343)	222 (348)
	Chinese	950K	750K	110K	90K	1,729 (2,280)	1,391 (1,810)	172 (252)	166 (218)
	Arabic	300K	240K	30K	30K	447 (447)	359 (359)	44 (44)	44 (44)

Table 3: Number of documents in the OntoNotes v5.0 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.

create the train/development/test partitions for English, Chinese and Arabic. We tried to reuse previously established partitions for Chinese and Arabic, but either they were not in the selection used for OntoNotes, or were partially overlapping, or had a very small portion of OntoNotes covered in the test set. Unfortunately, unlike English WSJ partitions, there was no clean way of reusing those partitions. Algorithm 1 details this procedure. The list of training/development/test document IDs can be found on the task webpage[12]. Following the recent CoNLL tradition, participants were allowed to use both the training and the development data to train their final model(s).

The number of documents in the corpus for this task, for each of the different languages, and for each of the training/development/test portions, are shown in Table 3. For comparison purposes, it 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 languages and genres similar to the ones in OntoNotes. In fact, there is some overlap between ACE and OntoNotes source documents.

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 was provided to the participants.

4.4.1 Manual Annotation *Gold* Layers

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

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 the ones in the broadcast conversation, weblogs, and telephone conversation genre — are very long and that prohibited efficient annotation in their entirety. These had to be split into smaller parts. A few passes to join some adjacent parts were conducted, 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 the purpose of 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 WalMart. The recent Treebank revision split tokens at *most* hyphens and 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. Unlike English, Chinese and Arabic have coreference annotation on elided subjects/objects. Recovering these entities in text is a hard problem, and the most recently reported numbers in literature for Chinese are around a F-score of 50 (Yang and Xue, 2010; Cai et al., 2011b). For Arabic there have not been much studies on recovering these. A study by Gabbard (2010) shows that these can be recovered with an F-score

[12] http://conll.cemantix.org/2012/download/ids/

For each language there are two sub-directories — "all" contains more general lists which include documents that had at least one of the layers of annotation, and "coref" contains the lists that include document that have coreference annotation. The former were used to generate training/development/test sets for layers other than coreference, and the latter was used to generate training/development/test sets for the coreference layer used in this shared task.

of 55 with automatic parses and roughly 65 using gold parses[13]. Considering the level of prediction accuracy of these tokens, and the relative frequency of the same, plus the fact that the CoNLL tabular format is not amenable to a variable number of tokens, we decided not to consider them as part of the task. In other words, we removed the manually identified traces (***pro*** and *) respectively in Chinese and Arabic Treebanks. We also do not consider the links that are formed by these tokens in the gold evaluation key.

Tables 4 and 5 shows the distribution of mentions by the syntactic categories, and the counts of entities, links and mentions in the corpus respectively. Interestingly the mentions formed by these dropped pronouns total roughly about 11% for both Chinese and Arabic. All of this data has been Treebanked and PropBanked either as part of the OntoNotes effort, or some previous effort.

Language	Syntactic	Train		Development		Test	
	category	Count	%	Count	%	Count	%
English	Noun Phrase	61.8K	39.46	9.7K	45.57	9.2K	42.97
	Pronoun	66.7K	42.61	7.8K	36.66	8.2K	38.69
	Proper Noun	18.1K	11.60	2.2K	10.66	2.3K	10.96
	Dropped Pro.	-	-	-	-	-	-
	Other Noun	2.636	1.68	546	2.55	500	2.33
	Verb	2.522	1.61	299	1.40	342	1.60
	Other	4.761	3.04	676	3.16	738	3.45
Chinese	Noun Phrase	40.7K	34.23	5.4K	32.53	5.1K	35.31
	Pronoun	20.8K	17.50	3.3K	19.88	2.5K	17.65
	Dropped Pro.	13.5K	11.39	1.9K	12.04	1.5K	10.71
	Proper Noun	19.0K	15.96	2.8K	17.24	2.2K	15.54
	Other Noun	23.6K	19.88	2.8K	17.08	2.8K	19.71
	Verb	244	0.20	51	0.31	20	0.14
	Other	994	0.83	153	0.92	139	0.95
Arabic	Noun Phrase	10.8K	34.93	1.3K	35.02	1.3K	36.51
	Pronoun	8.9K	28.77	1.0K	28.33	1.1K	30.58
	Dropped Pro.	3.5K	11.52	477	12.57	429	11.78
	Proper Noun	4.0K	13.01	450	11.86	390	10.71
	Other Noun	3.3K	10.90	439	11.57	345	9.47
	Verb	25	0.08	4	0.11	0	0.00
	Other	247	0.79	21	0.55	35	0.96

Table 4: Distribution of mentions in the data by their syntactic category.

Language	Type	Train	Development	Test	All
English	Entities/Chains	35,143	4,546	4,532	44,221
	Links	120,417	14,610	15,232	150,259
	Mentions	155,560	19,156	19,764	194,480
Chinese	Entities/Chains	28,257	3,875	3,559	35,691
	Links	74,597	10,308	9,242	94,147
	Mentions	102,854	14,183	12,801	129,838
Arabic	Entities/Chains	8,330	936	980	10,246
	Links	19,260	2,381	2,255	23,896
	Mentions	27,590	3,313	3,235	34,138

Table 5: Number of entities, links and mentions in the OntoNotes v5.0 data.

Parse Trees These represent the syntactic layer that is a revised version of the treebanks in English, Chinese and Arabic. Arabic treebank has probably seen the most revision over the past few years, in an effort to increase consistency. 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. As mentioned in the previous section, these include the cases of traces in Chinese and Arabic which are dropped subjects/objects that are legitimate targets for coreference annotation. 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 English parses of the OntoNotes, the disfluencies are marked using a special EDITED[14] 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,[15] 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 English Treebank, and the coreference chains were remapped to trees without disfluencies. Owing to various constraints, we decided to retain the disfluencies in the Chinese data. Since Arabic portion of the corpus is all newswire, this had no impact on it. However, for both Chinese and Arabic, since we remove trace tokens corresponding to dropped pronouns, all the other layers of annotation had to be remapped to the remaining sequence of tree tokens.

Propositions The propositions in OntoNotes are PropBank-style semantic roles for English, Chinese and Arabic. Most of the verb predicates in the corpus have been annotated with their arguments. As part of the OntoNotes effort, some enhancements were made to the English PropBank and Treebank to make them synchronize better with each other (Babko-Malaya et al., 2006). One of the outcomes of this effort was that two types of LINKs that represent pragmatic coreference (LINK-PCR) and selec-

[13] These numbers are not in the thesis, but we received them in an email communication with the Ryan Gabbard.

[14] 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 by different speakers, so we decided not to remove that from the data.

[15] 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.

tional preferences (LINK-SLC) were added to PropBank. More details can be found in the addendum to the PropBank guidelines[16] in the OntoNotes v5.0 release. 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 PropBank 1.0 release. This functionality is part of the OntoNotes DB Tool.[17]

Word Sense Gold standard word sense annotation was supplied using sense numbers (along with the sense inventories) as specified in the OntoNotes list of senses for each lemma. The coverage of the word sense annotation varies among the languages. English has the most coverage, while coverage for Chinese and Arabic is more sporadic. Even for English, the coverage for word sense annotation is not complete. Only some of the verbs and nouns are annotated with word sense information.

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 itself 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` file. There were some cases of interruptions and interjections that led to a sentence associated with 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 v5.0 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 v5.0 release.

Layer	English		Chinese	Arabic	
	Verb	Noun	All	Verb	Noun
Sense Inventories	2702	2194	763	150	111
Frames	5672	1335	20134	2743	532

Table 7: Number of senses defined for English, Chinese and Arabic in the OntoNotes v5.0 corpus.

Parse Trees Predicted parse trees for English were produced using the Charniak parser[18] (Charniak and Johnson, 2005). 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 extended correspondingly. Chinese and Arabic parses were generated using the Berkeley parser (Petrov and Klein, 2007). In the case of Arabic, the parsing community uses a mapping from rich Arabic part of speech tags, to Penn-style part of speech tags. We used the mapping that is included with the Arabic treebank.

The predicted parses for the training portion of the data were generated using 10-fold (5-fold for Arabic) cross-validation. The development and test parses were generated using a model trained on the entire training portion. We used OntoNotes v5.0 training data for training the Chinese and Arabic parser models, but the OntoNotes v4.0 subset of OntoNotes v5.0 data was used for training the English model. We decided to do the latter to be able to better compare the scores to the CoNLL-2011 evaluation given that parser is a central component to a coreference system, and the fact that OntoNotes v5.0 adds a small fraction of gold parses on top of those provided by OntoNotes v4.0. Table 6 shows the performance of the re-trained parsers on the CoNLL-2012 test set. We did not get a chance to re-train the re-ranker available for English, 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. In addition to the parser scores and part of speech accuracy, we have also added a column for the accuracy for the NPs because they are particularly relevant to the coreference task.

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

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

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

[19] There was an error in processing the test set, therefore the performance on the test set was slightly lower than the correct one reported in the table. The performance of the sense tagging the offical test set is 77.6 (R), 71.5 (P) and 74.4 (F).

		All Sentences						Sentence length < 40			
		N	POS	NP	R	P	F	N	R	P	F
English	Broadcast Conversation [BC]	2,194	95.93	90.05	84.30	84.46	84.38	2,124	85.83	85.97	85.90
	Broadcast News [BN]	1,344	96.50	91.11	84.19	84.28	84.24	1,278	85.93	86.04	85.98
	Magazine [MZ]	780	95.14	91.63	87.11	87.46	87.28	736	87.71	88.04	87.87
	Newswire [NW]	2,273	96.95	90.14	87.05	87.45	87.25	2,082	88.95	89.27	89.11
	Telephone Conversation [TC	1,366	93.52	88.96	79.73	80.83	80.28	1,359	79.88	80.98	80.43
	Weblogs and Newsgroups [WB]	1,658	94.67	89.16	83.32	83.20	83.26	1,566	85.14	85.07	85.11
	Pivot Text [PT] (New Testament)	1,217	96.87	95.39	92.48	93.66	93.07	1,217	92.48	93.66	93.07
	Overall	9,615	96.03	90.78	85.25	85.43	85.34	9145	86.86	87.02	86.94
Chinese	Broadcast Conversation [BC]	885	94.79	86.32	79.35	80.17	79.76	824	80.92	81.86	81.38
	Broadcast News [BN]	929	93.85	86.00	80.13	83.49	81.78	756	81.82	84.65	83.21
	Magazine [MZ]	451	97.06	92.40	83.85	88.48	86.10	326	85.64	89.80	87.67
	Newswire [NW]	481	94.07	79.70	77.28	82.26	79.69	406	79.06	83.84	81.38
	Telephone Conversation [TC]	968	92.22	80.15	69.19	71.90	70.52	942	69.59	72.24	70.89
	Weblogs and Newsgroups [WB]	758	92.37	85.60	78.92	82.57	80.70	725	79.30	83.10	81.16
	Overall	4,472	94.12	85.74	78.93	82.23	80.55	3,979	79.80	82.79	81.27
Arabic	Newswire [NW]	1,003	94.12	80.70	75.67	74.71	75.19	766	77.44	74.99	76.19

Table 6: Parser performance on the CoNLL-2012 test set.

		Accuracy		
		R	P	F
English	Broadcast Conversation [BC]	81.3	81.2	81.2
	Broadcast News [BN]	81.5	82.0	81.7
	Magazine [MZ]	78.8	79.1	79.0
	Newswire [NW]	85.7	85.7	85.7
	Weblogs and Newsgroups [WB]	77.6	77.5	77.5
	Overall	82.5	82.5	82.5
Chinese	Broadcast Conversation [BC]	-	-	80.5
	Broadcast News [BN]	-	-	85.4
	Magazine [MZ]	-	-	82.4
	Newswire [NW]	-	-	89.1
	Overall	-	-	84.3
Arabic	Newswire [NW][19]	75.2	75.9	75.6

Table 8: Word sense performance over both verbs and nouns in the CoNLL-2012 test set.

Word Sense This year we used the IMS (It Makes Sense) (Zhong and Ng, 2010) word sense tagger.[20] Word sense information, unlike syntactic parse information is not central to approaches taken by current coreference systems and so we decided to use a better word sense tagger to get a good state of the art accuracy estimate, at the cost of a completely fair (but, still close enough) comparison with English CoNLL-2011 results. This will also allow potential future uses to benefit from it. IMS was trained on all the word sense data that is present in the training portion of the OntoNotes corpus using cross-validated predictions on the input layers similar to the proposition tagger. During testing, for English and Arabic, IMS must first uses the automatic POS information to identify the nouns and verbs in the test data, and then assign senses to the automatically identified nouns and verbs. In case of Arabic, IMS uses gold lemmas. Since automatic POS tagging is not perfect, IMS does not always output a sense to all word tokens that need to be sense tagged due to wrongly predicted POS tags. As such, recall is not the same as precision on the English and Arabic test data. Recall that in Chinese, the word senses are defined against *lemmas* and are independent of the part of speech. Since we provide gold word segmentation, IMS attempts to sense tag all correctly segmented Chinese words, so recall and precision are same and so is F_1. Table 7 gives the number of lemmas covered by the word sense inventory in the English, Chinese and Arabic portion of OntoNotes.

Table 8 shows the performance of this classifier aggregated over *both the verbs and nouns* in the CoNLL-2012 test set. For English, genres PT and TC, and for Chinese genres TC and WB, no gold standard senses were available, and so their accuracies could not be computed.

[20] We offer special thanks to Hwee Tou Ng and his student Zhi Zhong for training IMS models and providing output for the development and test sets.

Propositions We used ASSERT[21] (Pradhan et al., 2005) to predict the propositional structure for English. Similar to the parser model for English, the same proposition model that was used in the CoNLL-2011 shared task — trained on all the training portion of the OntoNotes v4.0 data using cross-validated predicted parses — was used to generate the propositions for the development and test sets for this evaluation. We took a two stage approach to tagging where The NULL arguments are first filtered out, and the remaining NON-NULL arguments are classified into one of the argument types. The argument identification module used an ensemble of ten classifiers — each trained on a tenth of the training data and combined using unweighted voting. 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 (Pradhan et al., 2005). The Chinese propositional structure was predicted with the Chinese semantic role labeler described in (Xue, 2008), retrained on all the training portion of the OntoNotes v5.0 data. No propositional structures were provided for Arabic due to resource constraints. Table 9 shows the detailed performance numbers. The CoNLL-2005 scorer was used to compute the scores. At first glance, the performance on the English newswire genre is much lower than what has been reported for WSJ Section 23. This could be attributed to several factors: i) the fact that we had to compromise on the training method, ii) the newswire in OntoNotes not only contains WSJ data, but also Xinhua news, iii) The WSJ training and test portions in OntoNotes are a subset of the standard ones that have been used to report performance earlier; iv) the PropBank guidelines were significantly revised during the OntoNotes project in order to synchronize well with the Treebank, and finally v) it includes propositions for *be* verbs missing from the original PropBank. It looks like the newly added Pivot Text data (comprising of the New Testament) shows very good performance. This is not surprising given a similar trend in it parsing performance.

Framesets	Lemmas
1	2,722
2	321
> 2	181

Table 10: Frameset polysemy across lemmas.

In addition to automatically predicting the arguments, we also trained a classifier to tag PropBank frameset IDs for the English data. Table 7 lists the number of framesets available across the three languages[22] An overwhelming number of them are monosemous, but the more frequent verbs tend to be polysemous. Table 10 gives the distribution of number of framesets per lemma in the PropBank layer of the English OntoNotes v5.0 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.

Named Entities BBN's IdentiFinder^{TM}system was used to predict the named entities. For the CoNLL-2011 shared task we did not get a chance to re-train Identifinder, and used the stock model which did not have the same set of named entities as in the OntoNotes corpus, so we decided to

		Frameset Accuracy	Total Sentences	Total Propositions	% Perfect Propositions	Argument ID + Class		
						P	R	F
English	Broadcast Conversation [BC]	92	2,037	5,021	52.18	82.55	64.84	72.63
	Broadcast News [BN]	91	1,252	3,310	53.66	81.64	64.46	72.04
	Magazine [MZ]	89	780	2,373	47.16	79.98	61.66	69.64
	Newswire [NW]	93	1,898	4,758	39.72	80.53	62.68	70.49
	Telephone Conversation [TC]	90	1,366	1,725	45.28	79.60	63.41	70.59
	Weblogs and Newsgroups [WB]	92	929	2,174	39.19	81.01	60.65	69.37
	Pivot Corpus [PT]	92	1,217	2,853	50.54	86.40	72.61	78.91
	Overall	91	9,479	24,668	44.69	81.47	61.56	70.13
Chinese	Broadcast Conversation [BC]	-	885	2,323	31.34	53.92	68.60	60.38
	Broadcast News [BN]	-	929	4,419	35.44	64.34	66.05	65.18
	Magazine [MZ]	-	451	2,620	31.68	65.04	65.40	65.22
	Newswire [NW]	-	481	2,210	27.33	69.28	55.74	61.78
	Telephone Conversation [TC]	-	968	1,622	32.74	48.70	59.12	53.41
	Weblogs and Newsgroups [WB]	-	758	1,761	35.21	62.35	68.87	65.45
	Overall	-	4,472	14,955	32.62	61.26	64.48	62.83

Table 9: Performance on the propositions and framesets in the CoNLL-2012 test set.

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

[22] The number of lemmas for English in Table 10 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.

		All Genre	BC	BN	MZ	NW	TC	WB
		F	F	F	F	F	F	F
English	Cardinal	68.76	58.52	75.34	72.57	83.62	32.26	57.14
	Date	78.60	73.46	80.61	71.60	84.12	63.89	65.48
	Event	44.63	30.77	50.00	36.36	50.00	0.00	66.67
	Facility	47.29	64.20	43.14	40.00	54.17	0.00	28.57
	GPE	89.77	89.40	93.83	92.87	92.56	81.19	91.36
	Language	47.06	-	75.00	50.00	33.33	22.22	66.67
	Law	48.00	0.00	100.00	0.00	50.98	0.00	100.00
	Location	59.00	54.55	61.36	54.84	67.10	-	44.44
	Money	75.45	33.33	63.64	77.78	79.12	92.31	58.18
	NORP	88.58	94.55	93.92	94.87	90.70	78.05	85.15
	Ordinal	71.39	74.16	80.49	79.07	74.34	84.21	55.17
	Organization	76.00	60.90	78.57	69.97	84.76	48.98	51.08
	Percent	89.11	100.00	83.33	75.00	91.41	83.33	72.73
	Person	78.75	93.35	94.36	87.47	85.80	73.39	76.49
	Product	52.76	0.00	77.65	0.00	42.55	0.00	0.00
	Quantity	50.00	17.14	66.67	62.86	81.82	0.00	30.77
	Time	60.65	66.13	67.33	66.67	64.29	27.03	55.56
	Work of Art	34.03	42.42	35.62	28.57	54.24	0.00	8.70
	Overall	77.95	77.02	84.95	80.33	84.73	62.17	69.47

Table 11: Named Entity performance on the English subset of the CoNLL-2012 test set.

update the model for this round by retraining it on the English portion of the OntoNotes v5.0 corpus. Given the various constraints, we could not re-train it on the Chinese and Arabic data, Table 11 shows the overall performance of the tagger on the CoNLL-2012 English test set, as well as the performance broken down by individual name types.

Other Layers As noted earlier, systems were allowed to make use of gender and number predictions for NPs using the table from Bergsma and Lin (Bergsma and Lin, 2006), and the speaker metadata for broadcast conversations, telephone conversations and author or poster metadata for weblogs and newsgroups.

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 various annotation layers represented by file extensions for the documents at the leaves, and language, genre, source and section within a particular source forming the intermediate directories — `data/<language>/annotations/<genre>/<source>/<section>/<document>.<layer>`. It comes with various ways of querying and manipulating the data and allows convenient access to the information inside the sense inventory and PropBank frame files instead of having to interpret the raw `.xml`. 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 extension — `.conll` which logically served as another layer in addition to the `.parse`, `.prop`, `.sense`, `.name` and `.coref` layers which house the respective annotations. 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 we are 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 is essentially the `.conll` file, but with the column that contains the words, anonymized. 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[23]. Once the `.conll` file is created, it can be used to create the individual layers such as `.parse`, `.name`, and `.coref` that have inline annotation, with the provided scripts. We provide the layers that have standoff annotation (mostly with respect to the tokens in the treebank) like the `.prop` and `.sense` along with the `.skel` file.

In the CoNLL-2011 task, there were a few issues, where some teams used the test data accidentally during training. To prevent it from happening again

[23] 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 asterisk with the `([pos] [word])` string (or leaf) and concatenating the items in the rows of that column.
7	Lemma	The predicate/sense lemma is mentioned for the rows for which we have semantic role or word sense 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 4.
10	Speaker/Author	This is the speaker or author name where available. Mostly in Broadcast Conversation and Weblog 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 12: Format of the `.conll` file used in 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.

this year, we were advised by the steering committee to distribute the data in two installments. One for training and development and the other for testing. The test data released from LDC did not contain the coreference layer. Therefore, this year unlike previous CoNLL tasks, the test data contained some *truly unseen documents*. This made it easier to spot potential training errors such as ones that occurred in the CoNLL-2011 task. Table 12 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 for the shared task. Unlike propositions, word sense and named entities, where it is simply a matter of counting the correct answers, or for parsing, where there is an established metric, evaluating the accuracy of coreference continues to be contentious. Various alternative 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 makes the distinction between identity coreference and appositive coreference, treating the latter separately. 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-2012 coreference task are likely to be lower than for coreference evaluations based on MUC, or ACE data, where an approximate match is often allowed based on the specified head of the mentions.

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 that addresses all the concerns. Three metrics have been commonly used 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 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 is the oldest and most widely used. It focuses on the **links** (or, pairs of mentions) in the data.[24] 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 MUC'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 $\frac{|K \cap R|}{|K|}$ and precision for the same is is computed as $\frac{|K \cap R|}{|R|}$. 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$_m$. 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

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

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 after having consulted various researchers in the field, we came to a conclusion that each metric had its pros and cons and there is no silver bullet. Therefore we settled on the MELA metric proposed 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 is based on *links*. The B-CUBED is based on *mentions*, and the CEAF is based on *entities*. We decided to use the entity based CEAF_e instead of mention based CEAF_m. For a given end application, a weighted average of the three might be optimal, but since we don't have a particular 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. This still leaves us with a score for each language. We wanted to encourage researchers to run their systems on all three languages. Therefore, we decided to compute the final *official* score that would determine the winning submission as the average of the MELA metric across all the three languages. We decided to give a MELA score of zero to every language that a particular group did not run its system on.

4.5.3 Scoring Metrics Implementation

We used the same core scorer implementation[25] 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 since then.

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) have been 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-2012 task webpage.[26]

5 Participants

A total of 41 different groups demonstrated interest in the shared task by registering on the task webpage. Of these, 16 groups from 6 countries submitted system outputs on the test set during the evaluation week. 15 groups participated in at least one language in the closed task, and only one group participated solely in the open track. One participant (*yang*) did not submit a final task paper. Tables 13 and 14 list the distribution of the participants by country and the participation by language and task type.

Country	Participants
Brazil	1
China	8
Germany	3
Italy	1
Switzerland	1
USA	2

Table 13: Participation by country.

	Closed	Open	Combined
English	15	1	16
Chinese	13	3	14
Arabic	7	1	8

Table 14: Participation across languages and tracks.

6 Approaches

Tables 15 and 16 summarize the approaches taken by the participating systems along some important dimensions. While referring to the participating systems, as a convention, we will use the last name of the contact person from the participating team. It is almost always the last name of the first author of the system papers, or the first name in case of conflicting last names (*xinxin*). The only exception is *chunyang* which is the first name of the second author for that system. For space and readability purposes, while referring to the systems in the paper we will refer to the system by the primary contact name in italics instead of using explicit citations.

Most of the systems divided the problem into the typical two phases — first identifying the potential mentions in the text, and then linking the mentions to form coreference chains, or entities. Many systems used rule-based approaches for mention detection, though one, *yang* did use trained models, and *li* used a hybrid approach by adding mentions from a trained model to the ones identified using rules. All systems ran a post processing stage, after linking potential mentions together, to delete the remaining unlinked mentions. It was common for the systems to represent the markables (mentions) internally in

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

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

[27] The participant did not submit a final paper, so this information is based on an email correspondence.

Participant	Track	Languages	Syntax	Learning Framework	Markable Identification	Verb	Feature Selection	# Features	Train
fernandes	C	A, C, E	P	Latent Structure Perceptron	All noun phrases, pronouns and name entities	×	Latent feature induction and feature templates	196 templates (E); 197 (C) and 223 (A)	T + D
björkelund	C	A, C, E	P	LIBLINEAR for linking, and Maximum Entropy (Mallet) for anaphoricity	NP, PRP and PRP$ in all languages; PN and NR in Chinese; all NE in English. Classifier to exclude non-referential *pronouns* in English (with a probability of 0.95).	×	Greedy forward selection (semi-automatic)	28 feature templates (C) and 34 (E)[a]	T + D
chen	C, O	A, C, E	P	Hybrid — Sieve approach followed by language-specific heuristic pruning and language-independent learning based pruning; Genre specific models	NP, PRP and PRP$ and selected NE in English. NP and QP in Chinese. Exclude Chinese interrogative pronouns *what* and *where*. NP and selected NE in Arabic. Learning to prune non-referential mentions	×	Backward elimination	–	T
stamborg	C	A, C, E	D	Logistic Regression (LIBLINEAR)	NP, PRP and PRP$ in all languages; PN in Chinese; all NE in English. Exclude pleonastic *it* in English. Prune smaller mentions with same head.	×	Forward + Backward starting from CoNLL-2011 feature set for English and Soon feature set for Chinese and Arabic	18–37	T + D
martschat	C	A, C	D	Directed multigraph representation where the weights are learned over the training data (on top of BART (Versley et al., 2008))	Eight different mention types for English, and adjectival use for nations and a few NEs are filtered as well as embedded mentions and pleonastic pronouns. Four mention types in Chinese. Copulas are also handled appropriately.	×	×	In the form of negative and positive relations	20% of T (E); 15% of T (C)
chang	C	E, C	P	Latent Structure Learning	All noun phrases, pronouns and name entities	×	×	Chang, et al., 2011	T + D
uryupina	C	A, C, E	P	modification of BART using multi-objective optimization. Domain specific classifiers for *nw* and *bc* genre.	Standard rules for English and Classifier to identify markable NPs in Chinese and Arabic.	×	×	∼45	T + D
zhekova	C	A, C, E	P	Memory based learning (TIMBL)	NP, PRP and PRP$ in English, and all NP in Chinese and Arabic. Singleton classifier.	×	×	33	T + D
li	C	A, C, E	P	MaxEnt	All phrase types that are mentions in training are considered as mentions and a classifier is trained to identify potential mentions.	√	×	11 feature groups	T + D
yuan	C, O	E, C	P	C4.5 and deterministic rules	All noun phrases, pronouns and name entities	×	×	–	T + D
xu	C	E, C	P	Decision tree classifier and deterministic rules	All noun phrases, pronouns and selected named entities selected and overlapping mentions are considered when they are second-level NPs inside an NP, for example coordinating NPs	×	×	51 (E) and 61 (C)	T
chunyang	C	E, C	P	Rule-based (adaptation of Lee et al. 2011's sieve structure)	Chinese NP and pronouns using part of speech PN and names using part of speech NR excluding measure words and certain names	×	×	–	–
yang[27]	C	E	P	Structural SVM	Mention detection classifier	×	Same feature set, but per classifier	40	T (2011)
xinxin	C	E, C	P	MaxEnt	NP, PRP and PRP$ in English and Chinese	×	Greedy forward backward	71	T + D
shou	C	E	P	Modified version of Lee et al., 2011 sieve system					
xiong	O	A, C, E	P	Lee et al., 2011 system					

Table 15: 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. In the Train column T represents training data and D represents development data.

[a] Communication with Anders Björkelund.

Participant	Positive Training Examples	Negative Training Examples	Decoding
fernandes	Identify likely mentions with an aim to generate high recall using the sieve method proposed in (Lee et al., 2011). Create directed arcs between mention pairs using a set of rules		A constrained latent predictor finds the maximum scoring document tree among possible candidates
björkelund	Closest Antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	Stacked resolvers — i) Best first, ii) Pronoun closest first — closest first for pronouns and best first for other mentions and iii) cluster-mention; disallow transitive nesting; proper noun mentions processed first, followed by other nouns and pronouns
chen	Rule-based sieve approach followed by heuristic and learning based pruning		
stamborg	Closest Antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	Chinese and Arabic — Closest-first clustering for pronouns and Best-first clustering otherwise. English — closest-first for pronouns and averaged best-first clustering otherwise.
martschat	Weights are trained on part of the training data		Greedy clustering for English; Spectral clustering followed by greedy clustering for Chinese to reduce number of candidate antecedents.
chang	Closest Antecedent (Soon, 2001)	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 strategy; separate classifiers for pronominal and non-pronominal mentions for English. Single classifier for Chinese.
uryupina	Closest Antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	mention pair model without ranking as in Soon 2001
zhekova	Closest Antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	All definite phrases used to create a pair for each anaphor with each mention preceding it within a window of 10 (English, Chinese) or 7 (Arabic) sentences.
li	—		Best-first clustering
yuan	—		Deterministic NP-NP followed by PP-NP
xu	Window of sentences is used to determine positive and negative examples. For English a window of 5 sentences is used whereas for Chinese a window of 10 sentences is used		All-pair linking followed by pruning or correction using a set of rules for NE-NE and NP-NP mentions for sentences outside of a 5/10 sentence window in English and Chinese respectively
chunyang	Lee et al., 2011 system		
yang	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
xinxin	Closest Antecedent (Soon, 2001)	Negative examples in between anaphor and closest antecedent (Soon, 2001)	Best-first clustering method
shou	Modified version of Lee et al., 2011 coreference system		
xiong	Lee et al., 2011 system		

Table 16: Participating system profiles — Part II. This focuses on the way positive and negative examples were generated and the decoding strategy used.

terms of individual parse tree NP constituent spans. Some systems consider only mention-specific attributes while performing the clustering, but the recent trend seems to indicate a shared attribute model, where the attributes of an entity are determined collectively by heuristically merging the attribute types and values of its constituent mentions. For example, if a mention marked *singular* is clustered with another entity marked *plural*, then the collective number for the entity is assigned to be {*singular, plural*}. Various types of trained models were used for predicting coreference. For a learning-based system, generation of positive and negative examples is very important. The participating systems used a range of sentence windows surrounding the anaphor in generating these examples. 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 literature more recently. Following on the success of rule-based linking model in the CoNLL-2011 shared task, many systems used a completely rule-based linking model, or used it as a initializing, or intermediate step in a learning based system. A hybrid approach seems to be a central theme of many high scoring systems. Also, taking cue from last year's systems, almost all systems trained pleonastic *it* classifiers, and used speaker-based constraints/features for the conversation genre. Many systems used the predicted Arabic parts of speech that were mapped-down to Penn-style parts of speech, but *stamborg* used some heuristics to convert them back to the complex part of speech type, using more frequent mapping, to get better performance for Arabic. The *fernandes* system uses feature templates defined on mention pairs. *björkelund* mentions that disallowing transitive closures gave performance improvement of 0.6 and 0.4 respectively for English and Chinese/Arabic. *björkelund* also mentions seeing a considerable increase in performance after adding features that correspond to the Shortest Edit Script (Myers, 1986) between surface forms and unvocalised Buckwalter forms, respectively. These could be better at capturing the differences in gender and number signaled by certain morphemes than hand-crafted rules. *chen* built upon the sieve architecture proposed in Raghunathan et al. (2010) and added one more sieve — head match — for Chinese and modified two sieves. Some participants tried to incorporate peculiarities of the corpus in their systems. For example, *martschat* excluded adjectival nation names. Unlike English, and especially in absence of an external resource, it is hard to make a gender distinction in Arabic and Chinese. *martschat* used the information that 先生(sir) and 女士(lady) often suggest gender information. *bo* and *martschat* used plurality markers 们 to identify plurals. For example, 同学 (student) is singular and 同学们 (students) is plural. *bo* also uses a heuristic that if the word 和 (and) appears in the middle of a mention M, and the two parts separated by 和 are sub-mentions of M, then mention M is considered to be plural. Other words which have the similar meaning of 和, such as 同, 与 and 跟, are also considered. *uryupina* used the rich part of speech tags to classify pronouns into subtype, person number and gender. Chinese and Arabic do not have definite noun phrase markers like *the* in English. In contrast to English there is no strict enforcement of using definite noun phrases when referring to an antecedent in Chinese. Both 这次演说 (the talk) and 演说 (talk) can corefer with the antecedent 克林顿在河内大选的演说 (Clinton's talk during Hanoi election). This makes it very difficult to distinguish generic expressions from referential ones. *martschat* checks whether the phrase starts with a definite/demonstrative indicator (e.g., 这(this) or 那(that)) in order to identify demonstrative and definite noun phrases. For Arabic, *uryupina* considers as definite all mentions with definite head nouns (prefixed with "Al") and all the idafa constructs with a definite modifier. *chang* uses training data to identify inappropriate mention boundaries. They perform a relaxed matching between predicted mentions and gold mentions ignoring punctuation marks and mentions that start with one of the following: *adverb*, *verb*, *determiner*, and *cardinal number*. In another extreme, *xiong* translated Chinese and Arabic to English, and ran an English system and projected mentions back to the source languages. Unfortunately, it did not work quite well by itself. One issue that they faced was that many instances of pronouns did not have a corresponding mention in the source language (since we do not consider mentions formed by dropped subjects/objects). Nevertheless, using this in addition to performing coreference resolution in these languages could be useful. Similar to last year, most participants appear not to have focused much on eventive coreference, those coreference chains that build off verbs in the data. This usually means that nominal mentions that should have linked to the eventive verb were instead linked in with some other entity, or remained unlinked. 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 90% of mentions in the data are NPs and pronouns). Many of the trained systems were also able to improve their performance by using feature selection, the details varied depending on the example selection strategy and the classifier used.

7 Results

In this section we will take a look at the performance overview of various systems and then look at the performance for each language in various settings separately. For the official test, beyond the raw source text, coreference systems were provided only with the predictions for the other annotation layers (parses, semantic roles, word senses, and named entities). A high-level summary of the results for the systems on the primary evaluation for both *open* and *closed* tracks is shown in Table 17. The scores under the columns for each language are the average of MUC, BCUBED and CEAF$_e$ for that language. The column **Official Score** is the average of those per-language averages, but only for the **closed** track. If a participant did not participate in all three languages, then they got a score of zero for the languages that were not attempted. The systems are sorted in descending order of this final **Official Score**. The last two columns indicate whether the systems used only the training or both training and development for the final submissions. Most top performing systems used both training and development data for training the final system. Note that all the results reported here still used the same, *predicted* information for all input layers.

It can be seen that *fernandes* got the highest combined score (58.69) across all three languages and metrics. While scores for each individual language are lower than the figures cited for other corpora, it 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 combined score across all languages is purely for ranking purposes, and does not really tell much about each individual language. Owing to the ordering based on official score, not all the best performing systems for a particular language are in sequential order. Therefore, for easier reading, the scores of the top ranking system are in bold red, and the top four systems are underlined in the table.

Looking at the the English performance, we can see that *fernandes* gets the best average across the three selected metrics (MUC, BCUBED and CEAF$_e$). The next best system is *martschat* (61.31) followed very closely by *björkelund* (61.24) and then *chang* (60.18). The performance differences between the better-scoring systems were not large, with only about three points separating the top four systems, and only six out of a total of sixteen systems getting a score lower than 58 points which was the highest performing score in CoNLL-2011.[28]

In case of Chinese, it is seen that *chen* performs the best with a score of 62.24. This is then followed by *yuan* (60.69), and then *björkelund* (59.97) and *xu* (59.22). It is interesting to note that the scores for the top performing systems for both English and Chinese are very close. For all we know, this is just a coincidence. Also, for both English and Chinese, the top performing system is almost 2 points higher than the second best system.

On the Arabic language front, once again, *fernandes* has the highest score of 54.22, followed closely by *björkelund* (53.55) and then *uryupina* (50.41)

Since the majority of mentions in all the three languages are noun phrases or pronouns, the accuracy with which these are predicted in the parse trees should directly bear on the coreference scores. Since pronouns are a closed class and single words, the main focus falls on the accuracy of the noun phrases. By no means is the accuracy of noun phrases the only factor determining the overall coreference accuracy, but it cannot be ignored either. It can be observed that the coreference scores for the three languages are in the same trend as the noun phrase accuracies for those languages as seen in Table 6. Recall that in case of both Chinese and Arabic, there are roughly 11% instances of dropped pronouns that were not considered as part of the evaluation. The performance for Chinee and Arabic would decrease somewhat if these were considered in the set of gold mentions (and entities).

Tables 18 and 19 show similar information for the two supplementary tasks — one given *gold mention boundaries* (GB) and one given correct, *gold mentions* (GM). We have however, kept the same relative ordering of the system participants as in Table 17 for ease of reading. Looking at Table 18 carefully, we can see that for English and Arabic the relative ranking of the systems remain almost the same, except for a few outliers: *chang* performs the best given *gold mentions* — by almost 7 points over the next best performing system. In the case of Chinese, *chen* performs almost 6 points better than the official performance given *gold boundaries*, and another 9 points given *gold mentions* and almost 8 points better than the next best system using *gold mentions*. We will look at more details in the following sections.

As mentioned earlier in Section 4.2 we conducted some supplementary evaluations. These can be categorized by a combination of two parameters. One of which applies to both training and test set, and one can only apply to the test set. The two parameters are: i) Syntax and ii) Mention Quality. Syntax can take two values: i) *predicted* (PS), or ii) *gold* (GS), and can be applicable during either training or test; and, the mention quality can be of three values: i) No boundaries (NB), ii) Gold mention boundaries

[28] More precise comparison later in Section 8.

Participant	Open			Closed			Official	Final model	
	English	Chinese	Arabic	English	Chinese	Arabic	Score	Train	Dev
fernandes				**63.37**	58.49	**54.22**	**58.69**	√	√
björkelund				61.24	59.97	53.55	58.25	√	√
chen		**63.53**		59.69	**62.24**	47.13	56.35	√	×
stamborg				59.36	56.85	49.43	55.21	√	√
uryupina				56.12	53.87	50.41	53.47	√	√
zhekova				48.70	44.53	40.57	44.60	√	√
li				45.85	46.27	33.53	41.88	√	√
yuan		61.02		58.68	60.69		39.79	√	√
xu				57.49	59.22		38.90	√	×
martschat				61.31	53.15		38.15	√	×
chunyang				59.24	51.83		37.02	–	–
yang				55.29			18.43	√	×
chang				60.18	45.71		35.30	√	×
xinxin				48.77	51.76		33.51	√	√
shou				58.25			19.42	√	×
xiong	59.23	44.35	44.37				0.00	√	√

Table 17: Performance on primary **open** and **closed** tracks using all predicted information.

Participant	Open			Closed			Suppl.	Final model	
	English	Chinese	Arabic	English	Chinese	Arabic	Score	Train	Dev
fernandes				**63.16**	61.48	**53.90**	**59.51**	√	√
björkelund				60.75	62.76	53.50	59.00	√	√
chen		**70.00**		60.33	**68.55**	47.27	58.72	√	×
stamborg				57.35	54.30	49.59	53.75	√	√
zhekova				49.30	44.93	40.24	44.82	√	√
li				43.04	43.28	31.46	39.26	√	√
yuan				59.50	64.42		41.31	√	√
xu				56.47	64.08		40.18	√	×
chang				60.89			20.30	√	√

Table 18: Performance on supplementary **open** and **closed** tracks using all predicted information, given **gold mention boundaries**.

Participant	Open			Closed			Suppl.	Final model	
	English	Chinese	Arabic	English	Chinese	Arabic	Score	Train	Dev
fernandes				69.35	66.36	**63.49**	66.40	√	√
björkelund				68.20	69.92	59.14	65.75	√	√
chen		**78.98**		70.46	**77.77**	52.26	**66.83**	√	×
stamborg				68.66	66.97	53.35	62.99	√	√
zhekova				59.06	51.44	55.72	55.41	√	√
li				51.40	59.93	40.62	50.65	√	√
yuan				69.88	76.05		48.64	√	√
xu				63.46	69.79		44.42	√	×
chang				**77.22**			25.74	√	√

Table 19: Performance on supplementary **open** and **closed** tracks using all predicted information, given **gold mentions**.

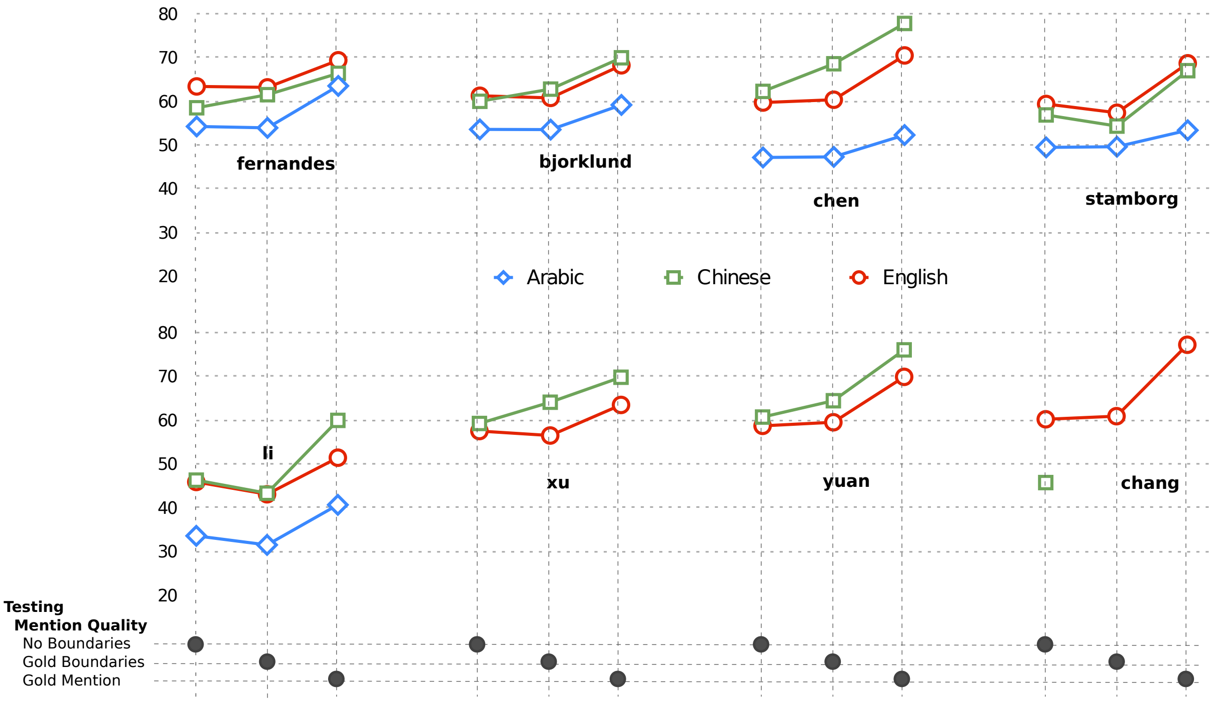

Figure 4: Performance for eight participating systems for the three languages, across the three mention qualities.

(GB) and iii) Gold mentions (GM), and can only be applicable during testing (since this information is not optional during training, as is the case with using gold or predicted syntax). There are a total of twelve combinations that we can form of using these parameters. Out of these, we thought six were particularly interesting. This is the product of the three cases of mention quality — NB, GB and GM, and two cases of syntax – GS and PS used during testing.

Figure 4 shows a performance plot for eight participating systems that attempted both the supplementary tasks — GB and GM in addition to the main NB for at least one of the three languages. These are all in the *closed* setting. At the bottom of the plot you can see dots that indicate what test condition to which a particular point refers. In most cases, for the hardest task — NB — the English and Chinese performances track quite close to each other. When provided with gold mention boundaries (GB), systems, *chen*, *xu* and *yuan* do significantly better in Chinese. There is almost no positive effect on the English performance across the board. In fact, performance of the *stamborg* and *li* systems drops noticeably. There is also a drop in performance for the *björkelund* system, but the difference is probably not significant. Finally, when provided with *gold mentions*, the performance of all systems increases across all languages, with *chang* showing the highest gain for English, and *chen* showing the highest gain for Chinese.

Figure 5 is a box and whiskers plot of the performance for all the systems for each language and variations — NB, GB, and GM. The circle in the center indicates the mean of the performances. The horizontal line in between the box indicates the median, and the bottom and top of the boxes indicate the first and third quartiles respectively, with the whiskers indicating the highest and lowest performance on that task. It can be easily seen that the English systems have the least divergence, with the divergence large for the GM case probably owing to *chang*. This is somewhat expected as this is the second year for the English task, and so it does show a more mature and stable performance. On the other hand, both Chinese and Arabic plots show much more divergence, with the Chinese and Arabic GB case showing the highest divergence. Also, except for Chinese GM condition, there is some skewness in the score distribution one way or the other.

Some participants ran their systems on six of the twelve possible combinations for all three languages. Figure 6 shows a plot for these tree participants — *fernandes*, *björkelund*, and *chen*. As in Figure 4, the dots at the bottom help identify which particular combination of parameters the point on the plot represents. In addition to the three test conditions related to mention quality, we now also have two more test conditions relating to the syntax.

We can see that the *fernandes* and *björkelund*, system performance tracks very close to each other. In other words, using gold standard parses during testing does not show much benefit in those cases. In case of *chen*, however, using gold parses shows a significant jump in scores for the NB condition. It seems that somehow, *chen* makes much better use of the gold parses. In fact, the performance is very close to the one with the GB condition. It is not clear what this system is doing differently that makes this possible. Adding more information, i.e., the GM condition, improves the performance by almost the same delta as going from NB to GB.

Finally, Figure 7 shows the plot for one system — *björkelund* — that was ran on ten of the twelve different settings. As usual the dots at the bottom help identify the conditions for a point on the plot. Now, there is a condition related to the quality of syntax during training as well. For some reasons, using *gold* syntax hurts performance — though slightly — in the NB and GB settings. Chinese does show some improvement when *gold* parse is used for training, only when *gold mentions* are available during testing.

One point to note is that we cannot compare these results to the ones obtained in the SEMEVAL-2010 coreference task which used a small portion of OntoNotes data because it was only using nominal entities, and had heuristically added singleton mentions[29].

[29] 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."

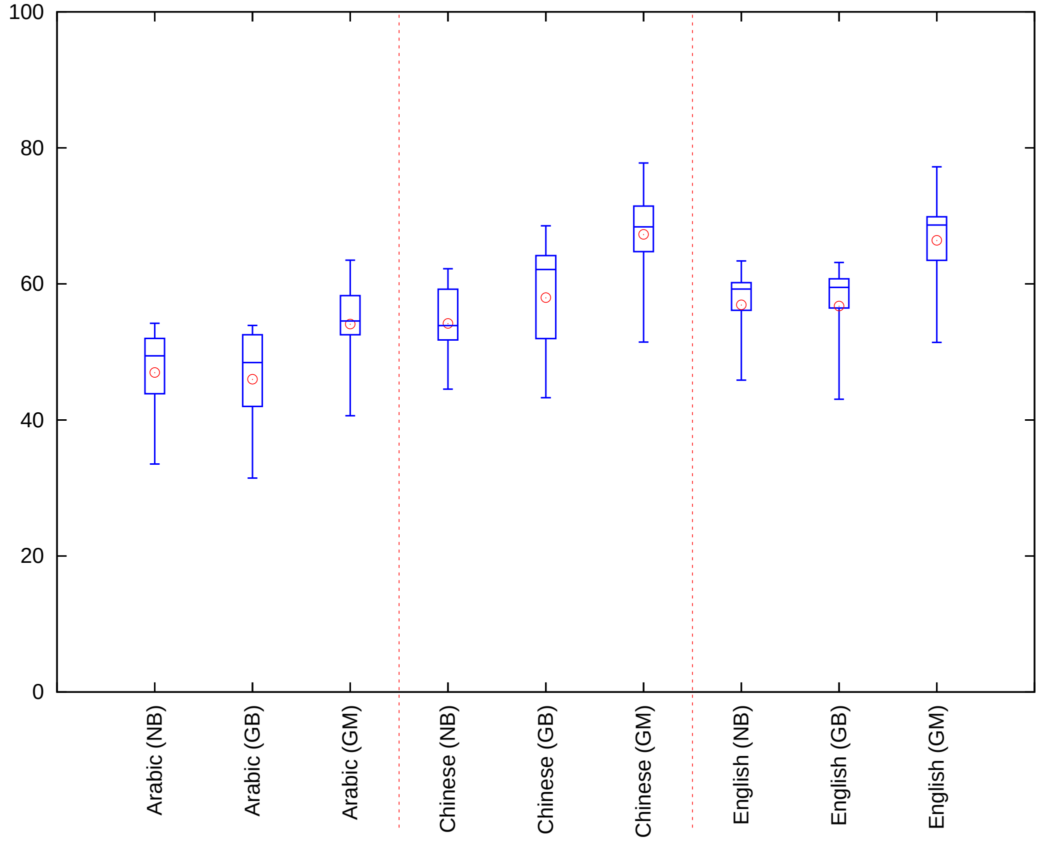

Figure 5: A box and whiskers plot of the performance for the three languages across the three mention qualities.

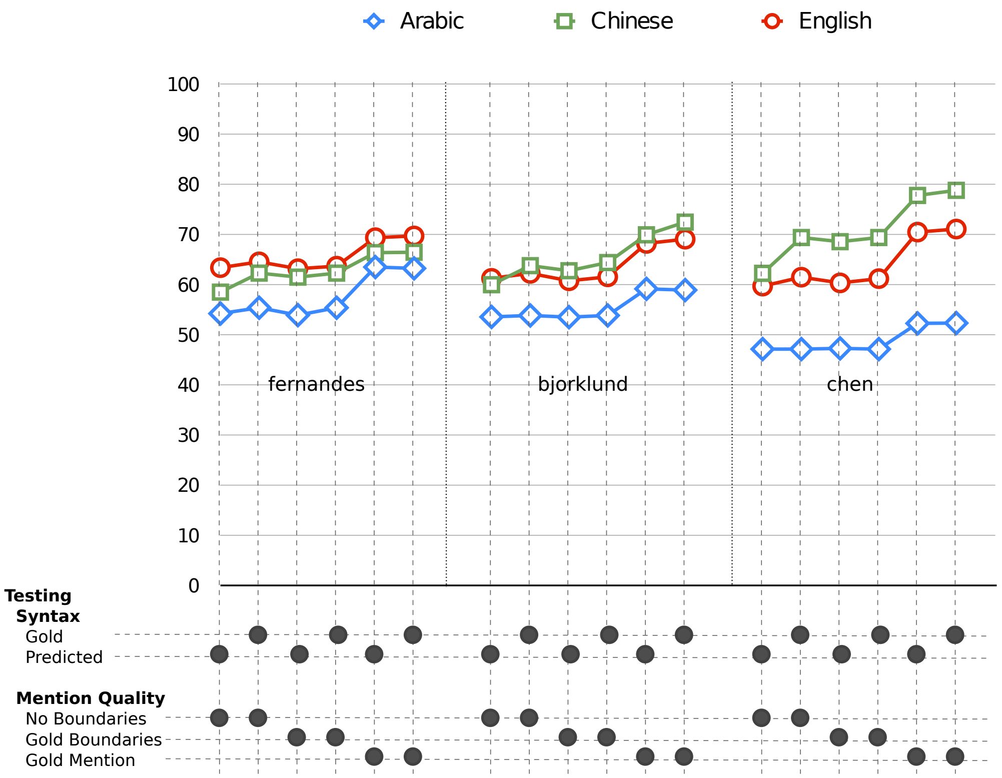

Figure 6: Performance of *fernandes*, *björkelund* and *chen* over six different settings.

In the following sections we will look at the results for the three languages, in various settings in more detail. It might help to describe the format of the tables first. Given that our choice of the official metric was somewhat arbitrary, it is also useful to look at the individual metrics. The tables are similar in structure to Table 20. Each table provides results across multiple dimensions. For completeness, the tables include the raw precision and recall scores from which the F-scores were derived. Each table shows the scores for a particular system for the task of *mention detection* and *coreference resolution* separately. The tables also include two additional scores (BLANC and CEAF$_m$) that did not factor into the official score. Useful further analysis may be possible based on these results beyond the preliminary results presented here. As you recall, OntoNotes does not contain any *singleton* mentions. Owing to this peculiar nature of the data, the mention detection scores cannot be interpreted independently of the coreference resolution scores. In this scenario, a mention is effectively an anaphoric mention that has at least one other mention coreferent with it in the document. 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 incorrectly 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. The tables are sub-divided into several logical horizontal sections separated by two horizontal lines. There can be a total of 12 sections, each categorized by a combination of two parse quality features GS and PS for each training and test set and three variations on the mention qualities — NB, GB, and GM, as described earlier. Just like we used the dots below the graphs earlier to indicate the parameters that were chosen for a particular point on the plot, we use small black squares in the tables after the participant name, to indicate the conditions chosen for the results on that particular row. Since there are many rows to each table, in order to facilitate finding which number we are referring to, we have added a ID column which uses letters **e**, **c**, and **a** to refer to the three languages — English, Chinese and Arabic. This is followed by a decimal number, in which the number before the decimal identifies the logical block within the table

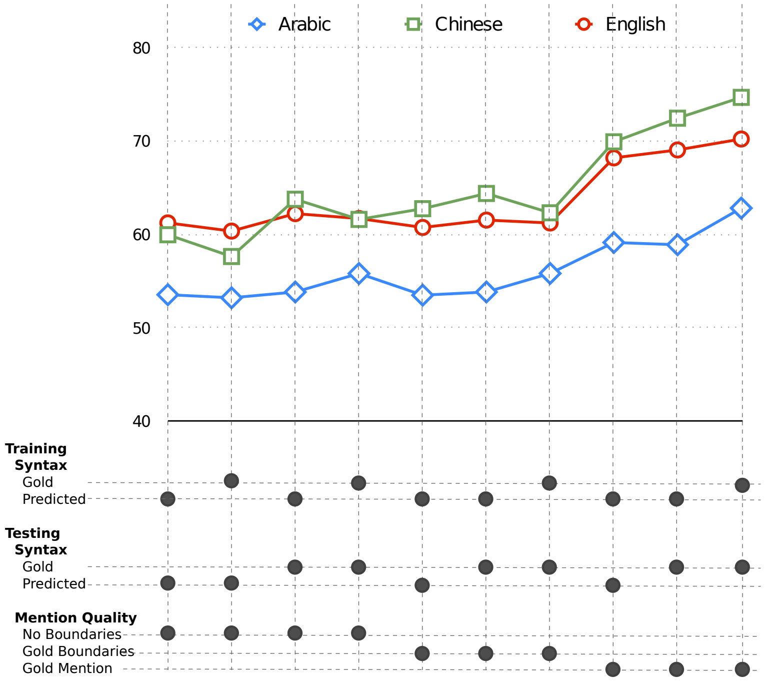

Figure 7: Performance of *björkelund* over ten different settings.

that share the same experiment parameters, and the one after the decimal indicates the index of a particular system in that block. Systems are sorted by the official score within each block. All the systems with NB setting are listed first, followed by GB, followed by GM. One participant (*björkelund*) ran more variations than we had originally planned, but since it falls under the general permutation and combination of the settings that we were considering, it makes sense to list those results here as well.

7.1 English *Closed*

Table 20 shows the performance for the English language in greater detail.

Official Setting Recall is quite important in the mention detection stage because the full coreference system has no way to recover if the mention detection stage misses a potentially anaphoric mention. The linking stage indirectly impacts the final mention detection accuracy. After a complete pass through the system some correct mentions could remain unlinked with any other mentions and would be deleted thereby lowering recall. Most systems tend to get a close balance between recall and precision for the mention detection task. A few systems had a considerable gap between the final mention detection recall and precision (*fernandes, xu, yang, li* and *xinxin*). It is not clear why this might be the case. One commonality between the ones that had a much higher precision than recall was that they used machine learned classifiers for mention detection. This could be possible because any classifier that is trained will not normally contain singleton mentions (as none have been annotated in the data) unless one explicitly adds them to the set of training examples (which is not mentioned in any of the respective system papers). A hybrid rule-based and machine learned model (*fernandes*) performed the best. Apart from some local differences, the ranking for all the systems is roughly the same irrespective of which metric is chosen. The CEAF$_e$ measure seems to penalize systems more harshly than the other measures. If the CEAF$_e$ measure does indicate the accuracy of entities in the response, this suggests that *fernandes* is doing better on getting coherent entities than any other system.

Gold Mention Boundaries In this case, all possible mention boundaries are provided to the system. This is very similar to what annotators see when they annotate the documents. One difficulty with this supplementary evaluation is that these boundaries alone provide only very partial information. For the roughly 10 to 20% of mentions that the automatic parser did not correctly identify, while the systems knew the correct boundaries, they had no structural syntactic or semantic information, and they also had to further approximate the already heuristic head word identification. This incomplete data complicates the systems' task and also complicates interpretation of the results. While most 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.[30]

Gold Mentions Another supplementary condition that we explored was if the systems were supplied with the manually-annotated spans for *all* and *only* 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 will also include verb headwords that are linked to event NPs, and will not include singleton mentions, which do not end up as part of any chain. The latter constraint makes this test seem 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 and anaphoricity determination has on the overall task and what the results are if they are perfectly known. 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 many of the systems did not get a 100% recall. This could possibly be owing to unlinked singletons that were removed in post-processing. *chang* along with *fernandes* are the only systems that got a perfect 100% recall. The reason is most likely because they had a hard constraint to link all mentions with at least one other mention. *chang* (77.22 [e7.00]) stands out in that it has a 7 point lead on the next best system in this category. This indicates that the linking algorithm for this system is significantly superior than the other systems — especially since the performance of the only other system that gets 100% mention score (*fernandes*) is much lower (69.35 [e7.03])

Gold Test Parses Looking at Table 20 it can be seen that there is a slight increase (∼1 point) in performance across all the systems when gold parses across all settings — NB, GB, and GM. In the case of *björkelund* for the NB setting, the overall performance improves by a percent when using gold test parse during testing (61.24 [e0.02] vs 62.23 [e1.02]), but strangely if gold parses are used during training as well, the performance is slightly lower (61.71 [e3.00]), although this difference is probably not statistically significant.

7.2 Chinese *Closed*

Table 21 shows the performance for the Chinese language in greater detail.

7.2.1 Official Setting

In this case, it turns out that *chen* does about 2 points better than the next best system across all the metrics. We know that this system had some more Chinese-specific improvements. It is strange that *fernandes* has a much lower mention recall with a much higher precision as compared to *chen*. As far as the system descriptions go, both systems seem to have used the same set of mentions — except for *chen* including QP phrases and not considering interrogative pronouns. One thing we found about *chen* was that they dealt with nested NPs differently in case of the NW genre to achieve some performance improvement. This unfortunately seems to be addressing a quirk in the Chinese newswire data owing to a possible data inconsistency in the release.

7.2.2 Gold Mention Boundaries

Unlike English, just the addition of *gold mention boundaries* improves the performance of almost all systems significantly. The delta improvement for *fernandes* turns out to be small, but it does gain on the mention recall as compared to the NB case. It is not clear why this might be the case. One explanation could be that the parser performance for constituents that represent mentions — primarily NP might be significantly worse than that for English. The mention recall of all the systems is boosted by roughly 10%.

7.2.3 Gold Mentions

Providing *gold mention* information further significantly boosts all systems. More so is the case with *chen* [e8.00] which gains another 9 points over the *gold mention boundary* condition in spite of the fact that they don't have a perfect recall. On the other hand, *fernandes* gets a perfect mention recall and precision, but ends up getting a 11 point lower performance [c8.05] than *chen*. Another thing to note is that for the CEAF$_e$ metric, the incremental drop in performance from the best to the next best and so on, is substantial, with a difference of 17 points between *chen* and *fernandes*. It does seem that the *chen* and *yuan* algorithm for linking is much better than the others.

7.2.4 Gold Test Parses

When provided with *gold parses* for the test set, there is a substantial increase in performance for the NB condition – numerically more so than in case of English. The degree of improvement decreases for the GB and GM conditions.

[30] 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.

ID	Participant	Train		Test					MENTION DETECTION			COREFERENCE RESOLUTION															Official
		Syntax		Syntax		Mention Qlty.						MUC			BCUBED			CEAF$_m$			CEAF$_e$			BLANC			
		A	G	A	G	NB	GB	GM	R	P	F	R	P	F$_1$	R	P	F$_2$	R	P	F	R	P	F$_3$	R	P	F	$\frac{F_1+F_2+F_3}{3}$
e0.00	fernandes	■		■		■			72.75	83.45	77.73	65.83	75.91	70.51	65.79	77.69	71.24	61.94	61.94	61.94	55.00	43.17	48.37	77.35	78.07	77.70	63.37
e0.01	martschat	■		■		■			74.23	76.10	75.15	65.21	68.83	66.97	66.50	74.69	70.36	59.61	59.61	59.61	48.64	44.72	46.60	73.29	78.94	75.73	61.31
e0.02	björkelund	■		■		■			73.75	77.09	75.38	65.23	70.10	67.58	65.90	75.24	70.26	59.20	59.20	59.20	48.60	43.42	45.87	73.47	78.81	75.80	61.24
e0.03	chang	■		■		■			72.48	76.25	74.32	64.77	68.06	66.38	67.04	71.81	69.34	58.28	58.28	58.28	46.64	43.11	44.81	75.24	74.73	74.98	60.18
e0.04	chen	■		■		■			75.08	72.60	73.82	63.47	63.96	63.71	66.57	71.52	68.96	58.13	58.13	58.13	46.68	46.15	46.41	71.30	79.15	74.48	59.69
e0.05	stamborg	■		■		■			75.51	72.39	73.92	66.26	63.98	65.10	69.09	69.54	69.31	56.76	56.76	56.76	42.53	44.89	43.68	74.03	77.28	75.52	59.36
e0.06	chunyang	■		■		■			75.23	72.24	73.71	64.08	63.57	63.82	66.45	70.71	68.51	57.24	57.24	57.24	45.13	45.67	45.40	71.12	77.92	73.95	59.24
e0.07	yuan	■		■		■			73.19	71.88	72.53	62.08	63.02	62.55	66.23	70.45	68.27	57.28	57.28	57.28	45.74	44.74	45.23	72.05	76.86	74.16	58.68
e0.08	shou	■		■		■			75.35	72.08	73.68	63.46	62.39	62.92	65.31	68.90	67.05	55.68	55.68	55.68	44.20	45.35	44.77	69.43	75.08	71.81	58.25
e0.09	xu	■		■		■			62.64	84.55	71.96	59.11	75.18	66.18	62.28	73.29	67.34	53.19	53.19	53.19	48.29	32.64	38.95	75.16	68.37	70.95	57.49
e0.10	uryupina	■		■		■			71.82	69.96	70.88	61.00	60.78	60.89	63.59	68.48	65.95	52.44	52.44	52.44	41.42	41.64	41.53	67.40	72.83	69.65	56.12
e0.11	yang	■		■		■			64.28	73.99	68.79	55.29	65.20	59.84	60.38	72.74	65.99	51.90	51.90	51.90	45.20	35.95	40.05	68.69	71.66	70.03	55.29
e0.12	xinxin	■		■		■			73.93	54.55	62.78	55.48	42.72	48.27	66.93	56.67	61.37	44.83	44.83	44.83	31.68	43.55	36.68	64.77	66.14	65.42	48.77
e0.13	zhekova	■		■		■			65.78	68.49	67.11	54.28	52.79	53.52	62.26	54.90	58.35	43.54	43.54	43.54	33.52	34.96	34.22	67.23	58.77	60.63	48.70
e0.14	li	■		■		■			45.78	86.72	59.93	39.12	72.57	50.84	43.03	80.06	55.98	41.97	41.97	41.97	49.44	22.30	30.74	64.01	66.86	65.24	45.85
e1.00	fernandes	■			■	■			74.76	84.81	79.47	67.73	77.25	72.18	66.42	78.01	71.75	63.05	63.05	63.05	56.16	44.51	49.66	76.89	78.60	77.71	64.53
e1.01	martschat	■			■	■			76.87	77.33	77.10	67.90	70.37	69.11	67.83	75.34	71.39	60.92	60.92	60.92	49.38	46.61	47.96	74.07	79.77	76.54	62.82
e1.02	björkelund	■			■	■			75.53	77.86	76.68	67.00	71.17	69.02	66.56	75.71	70.84	59.90	59.90	59.90	49.22	44.68	46.84	73.42	80.19	76.27	62.23
e1.03	chen	■			■	■			77.13	75.15	76.13	66.30	66.99	66.65	67.73	72.77	70.16	59.70	59.70	59.70	47.99	47.21	47.60	72.63	80.08	75.70	61.47
e1.04	zhekova	■			■	■			66.05	69.62	67.79	54.45	53.59	54.02	61.66	55.62	58.48	43.71	43.71	43.71	33.82	34.65	34.23	66.84	59.27	61.20	48.91
e2.00	björkelund		■	■		■			76.23	72.64	74.39	66.50	65.22	65.85	68.45	71.01	69.71	58.35	58.35	58.35	44.85	46.21	45.52	74.36	77.29	75.72	60.36
e3.00	björkelund		■		■	■			78.68	73.83	76.18	68.96	66.73	67.83	69.70	71.47	70.58	59.55	59.55	59.55	45.55	47.97	46.73	74.98	77.92	76.35	61.71
e4.00	fernandes	■		■			■		71.91	85.29	78.03	64.92	77.53	70.67	64.25	78.95	70.85	61.59	61.59	61.59	56.48	41.69	47.97	76.28	77.87	77.05	63.16
e4.01	chang	■		■			■		72.01	79.83	75.72	64.58	71.36	67.80	66.07	73.87	69.75	58.51	58.51	58.51	49.14	41.71	45.12	75.92	73.95	74.88	60.89
e4.02	björkelund	■		■			■		71.95	78.97	75.30	63.44	71.63	67.29	63.95	76.59	69.70	58.48	58.48	58.48	50.00	41.35	45.27	73.05	78.99	75.59	60.75
e4.03	chen	■		■			■		74.78	75.70	75.24	63.26	66.78	64.97	65.38	73.56	69.23	58.57	58.57	58.57	48.81	44.92	46.79	71.85	80.22	75.20	60.33
e4.04	yuan	■		■			■		75.73	70.84	73.20	64.50	62.79	63.64	68.53	69.99	69.25	57.90	57.90	57.90	44.73	46.53	45.61	72.86	76.93	74.69	59.50
e4.05	stamborg	■		■			■		76.46	67.30	71.59	66.16	58.80	62.26	71.21	64.98	67.95	55.20	55.20	55.20	38.46	45.83	41.83	76.19	73.58	74.80	57.35
e4.06	xu	■		■			■		66.31	75.82	70.74	62.15	67.13	64.55	67.53	66.97	67.25	52.02	52.02	52.02	40.04	35.45	37.60	77.29	67.42	70.74	56.47
e4.07	zhekova	■		■			■		66.45	70.91	68.61	54.96	54.67	54.82	61.85	55.60	58.56	43.96	43.96	43.96	34.38	34.67	34.53	68.49	59.51	61.52	49.30
e4.08	li	■		■			■		60.00	44.47	51.08	44.17	33.66	38.21	66.37	53.93	59.51	39.30	39.30	39.30	27.53	36.51	31.39	63.26	60.00	61.33	43.04
e5.00	fernandes	■			■		■		72.69	86.11	78.83	65.65	78.26	71.40	64.36	79.09	70.97	62.00	62.00	62.00	57.36	42.23	48.65	75.89	78.28	77.02	63.67
e5.01	björkelund	■			■		■		73.24	79.25	76.13	64.75	72.29	68.31	64.62	77.01	70.27	59.20	59.20	59.20	50.45	42.36	46.05	73.22	80.28	76.17	61.54
e5.02	chen	■			■		■		75.59	76.58	76.08	64.67	68.07	66.33	66.09	74.02	69.83	59.38	59.38	59.38	49.36	45.55	47.38	72.09	80.41	75.43	61.18
e5.03	stamborg	■			■		■		77.17	67.14	71.81	66.88	58.82	62.59	71.55	64.97	68.10	55.30	55.30	55.30	38.28	46.34	41.93	76.16	73.62	74.81	57.54
e5.04	zhekova	■			■		■		65.82	71.72	68.65	54.68	55.51	55.09	61.22	56.59	58.82	43.90	43.90	43.90	34.85	34.04	34.44	68.10	59.76	61.79	49.45
e6.00	björkelund		■		■		■		76.26	75.47	75.86	66.55	68.00	67.27	67.60	73.05	70.22	59.04	59.04	59.04	46.99	45.42	46.19	74.60	78.38	76.32	61.23
e7.00	chang	■		■				■	100.00	100.00	100.00	83.16	88.48	85.74	75.36	79.69	77.46	73.76	73.76	73.76	75.38	62.71	68.46	81.23	78.99	80.05	77.22
e7.01	chen	■		■				■	80.82	100.00	89.39	72.29	89.40	79.94	64.60	85.92	73.75	68.35	68.35	68.35	76.25	46.40	57.69	75.41	84.80	79.12	70.46
e7.02	yuan	■		■				■	80.03	100.00	88.91	72.22	89.16	79.80	64.75	84.68	73.39	67.76	67.76	67.76	74.49	45.46	56.46	75.60	83.10	78.69	69.88
e7.03	fernandes	■		■				■	100.00	100.00	100.00	70.69	91.21	79.65	65.46	85.61	74.19	67.64	67.64	67.64	74.71	42.55	54.22	79.41	80.17	79.78	69.35
e7.04	stamborg	■		■				■	78.17	100.00	87.74	71.22	88.12	78.77	64.75	83.16	72.81	66.74	66.74	66.74	71.94	43.74	54.41	78.68	81.47	79.99	68.66
e7.05	björkelund	■		■				■	75.69	100.00	86.16	69.51	90.71	78.70	62.37	87.03	72.67	66.32	66.32	66.32	74.14	41.52	53.23	76.49	82.90	79.22	68.20
e7.06	xu	■		■				■	67.09	100.00	80.30	64.81	89.94	75.34	62.88	80.57	70.63	59.13	59.13	59.13	65.28	33.67	44.42	78.60	72.37	74.87	63.46
e7.07	zhekova	■		■				■	78.55	100.00	87.98	68.38	78.11	72.92	63.04	58.60	60.74	50.35	50.35	50.35	52.64	37.10	43.53	69.87	61.19	62.74	59.06
e7.08	li	■		■				■	57.18	99.99	72.75	48.04	81.51	60.45	44.13	81.21	57.18	47.82	47.82	47.82	61.82	25.98	36.58	65.26	69.93	67.12	51.40
e8.00	chen	■			■			■	81.46	100.00	89.78	73.22	89.69	80.62	65.32	85.87	74.20	68.87	68.87	68.87	76.43	47.26	58.40	75.66	84.84	79.31	71.07
e8.01	fernandes	■			■			■	100.00	100.00	100.00	71.18	91.24	79.97	65.81	85.51	74.38	67.83	67.83	67.83	74.93	43.09	54.72	79.65	80.31	79.97	69.69
e8.02	stamborg	■			■			■	78.83	100.00	88.16	71.96	88.40	79.33	65.31	83.28	73.21	67.26	67.26	67.26	72.29	44.48	55.07	78.92	81.59	80.17	69.20
e8.03	björkelund	■			■			■	76.94	100.00	86.97	70.66	91.15	79.60	62.90	87.56	73.21	66.90	66.90	66.90	74.88	42.65	54.35	76.42	84.43	79.71	69.05
e8.04	zhekova	■			■			■	78.73	100.00	88.10	68.54	78.10	73.01	63.14	58.63	60.80	50.61	50.61	50.61	52.84	37.44	43.83	70.28	61.55	63.23	59.21
e9.00	björkelund		■		■			■	79.55	100.00	88.61	72.59	90.04	80.38	64.94	85.72	73.89	68.14	68.14	68.14	74.77	45.27	56.40	78.23	83.31	80.48	70.22

Table 20: Performance of systems in the *primary* and *supplementary* evaluations for the *closed track* for English.

ID	Participant	Train		Test					MENTION DETECTION			COREFERENCE RESOLUTION															Official
		Syntax		Syntax		Mention Qlty.						MUC			BCUBED			CEAF$_m$			CEAF$_e$			BLANC			
		A	G	A	G	NB	GB	GM	R	P	F	R	P	F$_1$	R	P	F$_2$	R	P	F	R	P	F$_3$	R	P	F	$\frac{F_1+F_2+F_3}{3}$
c0.00	chen	■		■		■			71.12	72.16	71.64	59.92	64.69	62.21	69.73	77.81	73.55	62.18	62.18	62.18	53.43	48.73	50.97	72.79	84.53	77.34	62.24
c0.01	yuan	■		■		■			72.75	64.09	68.15	62.36	58.42	60.33	73.12	72.67	72.90	59.59	59.59	59.59	47.10	50.70	48.83	73.72	78.22	75.76	60.69
c0.02	björkelund	■		■		■			69.45	63.54	66.37	58.72	58.49	58.61	71.23	75.07	73.10	59.01	59.01	59.01	48.09	48.29	48.19	72.25	81.61	76.07	59.97
c0.03	xu	■		■		■			64.33	66.10	65.20	55.02	61.47	58.07	68.39	76.38	72.16	57.46	57.46	57.46	50.40	44.81	47.44	71.64	74.57	73.00	59.22
c0.04	fernandes	■		■		■			57.24	78.28	66.13	52.69	70.58	60.34	62.99	80.57	70.70	57.73	57.73	57.73	53.75	37.88	44.44	75.06	79.59	77.11	58.49
c0.05	stamborg	■		■		■			57.65	71.93	64.01	52.56	64.13	57.77	64.43	77.55	70.38	55.57	55.57	55.57	47.90	38.04	42.41	72.74	77.84	75.00	56.85
c0.06	uryupina	■		■		■			50.98	70.09	59.03	45.62	63.13	52.97	59.17	80.78	68.31	52.40	52.40	52.40	48.47	34.52	40.32	68.72	80.76	73.11	53.87
c0.07	martschat	■		■		■			48.49	74.02	58.60	42.71	67.80	52.41	55.37	85.24	67.13	51.30	51.30	51.30	51.81	32.46	39.92	63.96	82.81	69.18	53.15
c0.08	chunyang	■		■		■			61.11	62.12	61.61	50.02	49.64	49.83	65.81	65.50	65.66	49.88	49.88	49.88	39.84	40.17	40.00	67.12	65.83	66.45	51.83
c0.09	xinxin	■		■		■			55.68	56.09	55.89	47.64	48.55	48.09	66.16	70.60	68.31	49.92	49.92	49.92	39.26	38.53	38.89	69.48	73.91	71.44	51.76
c0.10	li	■		■		■			36.60	87.01	51.53	32.48	71.44	44.65	45.51	86.06	59.54	45.70	45.70	45.70	55.11	25.24	34.62	64.99	76.63	68.92	46.27
c0.11	chang	■		■		■			39.43	59.97	47.58	30.85	49.22	37.93	53.02	78.31	63.23	44.89	44.89	44.89	44.92	29.99	35.97	59.82	71.24	63.16	45.71
c0.12	zhekova	■		■		■			35.12	72.52	47.32	31.19	57.97	40.56	49.49	77.65	60.45	41.86	41.86	41.86	45.92	25.24	32.58	64.29	61.64	62.79	44.53
c1.00	chen	■			■	■			83.22	79.25	81.19	72.14	72.77	72.46	75.32	80.20	77.68	68.67	68.67	68.67	58.37	57.64	58.00	76.48	87.07	80.80	69.38
c1.01	yuan	■			■	■			83.69	70.33	76.43	73.73	65.50	69.38	77.97	74.37	76.13	64.77	64.77	64.77	49.99	58.38	53.86	77.12	79.84	78.41	66.46
c1.02	xu	■			■	■			75.02	73.07	74.03	65.95	69.28	67.58	73.28	77.65	75.40	62.53	62.53	62.53	53.90	50.72	52.26	76.50	75.17	75.81	65.08
c1.03	björkelund	■			■	■			77.77	68.03	72.57	66.76	63.36	65.01	74.21	75.84	75.02	62.17	62.17	62.17	49.74	52.92	51.28	74.38	82.36	77.77	63.77
c1.04	fernandes	■			■	■			63.83	81.73	71.68	59.35	74.49	66.07	66.31	81.43	73.10	61.19	61.19	61.19	55.97	41.50	47.66	78.11	81.28	79.60	62.28
c1.05	stamborg	■			■	■			63.07	75.21	68.61	58.32	67.95	62.76	67.47	78.54	72.58	58.98	58.98	58.98	49.80	41.15	45.06	76.29	80.09	78.05	60.13
c1.06	uryupina	■			■	■			56.61	71.38	63.14	50.74	64.53	56.81	61.78	80.11	69.76	54.59	54.59	54.59	48.94	37.49	42.45	70.37	80.87	74.42	56.34
c1.07	martschat	■			■	■			52.94	77.17	62.80	47.30	72.19	57.15	57.31	86.69	69.00	53.92	53.92	53.92	54.11	34.40	42.06	65.39	84.86	70.95	56.07
c1.08	chunyang	■			■	■			67.52	65.33	66.41	56.11	52.81	54.41	68.27	64.49	66.33	51.92	51.92	51.92	40.39	43.47	41.88	69.16	65.03	66.80	54.21
c1.09	xinxin	■			■	■			62.59	58.21	60.32	53.49	50.44	51.92	68.70	69.78	69.24	51.95	51.95	51.95	39.65	42.18	40.87	71.40	74.66	72.90	54.01
c1.10	yang	■			■	■			58.66	56.52	57.57	48.74	49.49	49.11	65.61	72.91	69.07	49.61	49.61	49.61	40.11	39.52	39.81	64.51	73.91	67.95	52.66
c1.11	li	■			■	■			40.34	89.74	55.66	34.85	73.77	47.33	46.00	86.06	59.95	46.79	46.79	46.79	56.58	25.77	35.42	66.01	77.39	69.96	47.57
c1.12	chang	■			■	■			43.28	62.16	51.03	33.36	50.29	40.11	53.96	77.60	63.65	45.94	45.94	45.94	45.06	31.10	36.80	60.40	71.57	63.79	46.85
c1.13	zhekova	■			■	■			37.84	74.84	50.27	33.95	60.29	43.44	50.95	77.28	61.41	43.34	43.34	43.34	46.68	26.13	33.50	65.98	62.15	63.73	46.12
c2.00	björkelund		■	■		■			74.86	55.07	63.46	61.06	48.92	54.32	76.08	67.35	71.45	57.39	57.39	57.39	42.39	53.10	47.15	73.42	77.96	75.48	57.64
c3.00	björkelund		■		■	■			85.29	59.87	70.36	70.92	54.07	61.36	79.51	67.83	73.21	60.80	60.80	60.80	43.61	59.31	50.26	76.26	78.94	77.53	61.61
c4.00	chen	■		■			■		81.99	78.97	80.45	70.76	72.12	71.43	74.37	79.91	77.04	67.87	67.87	67.87	57.95	56.41	57.17	75.95	86.75	80.32	68.55
c4.01	yuan	■		■			■		82.89	66.86	74.02	72.12	61.59	66.44	77.96	72.30	75.02	62.96	62.96	62.96	47.17	57.47	51.81	75.77	78.48	77.05	64.42
c4.02	xu	■		■			■		73.21	72.68	72.94	63.54	68.73	66.03	71.36	78.70	74.85	61.57	61.57	61.57	53.90	49.07	51.37	72.90	77.06	74.80	64.08
c4.03	björkelund	■		■			■		71.88	70.18	71.02	61.69	65.54	63.56	70.42	79.14	74.52	61.32	61.32	61.32	52.03	48.51	50.20	72.44	85.49	77.40	62.76
c4.04	fernandes	■		■			■		64.21	79.18	70.91	58.76	71.46	64.49	66.62	79.88	72.65	60.40	60.40	60.40	54.09	42.02	47.29	77.69	80.96	79.22	61.48
c4.05	stamborg	■		■			■		71.21	55.28	62.24	61.13	47.20	53.27	75.47	63.09	68.73	52.82	52.82	52.82	35.81	47.71	40.91	74.53	71.12	72.69	54.30
c4.06	zhekova	■		■			■		36.97	73.98	49.30	32.09	58.30	41.39	49.43	77.38	60.32	42.09	42.09	42.09	46.35	25.71	33.07	63.41	61.47	62.34	44.93
c4.07	li	■		■			■		68.22	41.88	51.90	52.56	30.62	38.70	76.15	48.52	59.27	41.06	41.06	41.06	25.14	43.51	31.86	67.82	58.70	61.47	43.28
c5.00	chen	■			■		■		83.22	79.25	81.19	72.14	72.77	72.46	75.32	80.20	77.68	68.67	68.67	68.67	58.37	57.64	58.00	76.48	87.07	80.80	69.38
c5.01	yuan	■			■		■		83.75	67.46	74.72	73.25	62.49	67.44	78.46	72.68	75.46	63.52	63.52	63.52	47.54	58.13	52.30	76.10	78.79	77.37	65.07
c5.02	björkelund	■			■		■		76.53	70.28	73.27	65.46	65.87	65.66	72.79	78.38	75.49	62.91	62.91	62.91	52.21	51.82	52.01	73.70	85.52	78.36	64.39
c5.03	fernandes	■			■		■		63.83	81.73	71.68	59.35	74.49	66.07	66.31	81.43	73.10	61.19	61.19	61.19	55.97	41.50	47.66	78.11	81.28	79.60	62.28
c5.04	stamborg	■			■		■		71.36	55.32	62.32	61.17	47.27	53.33	75.64	63.41	68.99	53.26	53.26	53.26	36.11	48.05	41.23	75.75	72.53	74.02	54.52
c5.05	zhekova	■			■		■		37.89	74.79	50.30	33.93	60.19	43.39	50.87	77.27	61.35	43.29	43.29	43.29	46.62	26.13	33.49	65.91	62.31	63.81	46.08
c5.06	li	■			■		■		74.10	40.23	52.14	56.15	28.50	37.81	79.00	43.67	56.25	39.32	39.32	39.32	22.49	45.72	30.15	68.12	57.08	59.83	41.40
c6.00	björkelund		■		■		■		84.09	61.19	70.83	69.85	55.71	61.98	78.57	69.93	74.00	61.52	61.52	61.52	45.23	58.43	50.99	75.38	81.33	78.02	62.32
c7.00	björkelund		■		■			■	81.07	100.00	89.55	72.72	92.28	81.34	69.47	91.38	78.93	72.24	72.24	72.24	81.32	52.45	63.77	77.30	89.40	82.09	74.68
c8.00	chen	■		■				■	84.72	100.00	91.73	76.60	92.41	83.77	72.95	91.43	81.15	75.83	75.83	75.83	83.56	57.86	68.38	79.21	91.38	84.09	77.77
c8.01	yuan	■		■				■	81.72	99.79	89.85	74.77	92.74	82.79	70.91	91.21	79.79	73.67	73.67	73.67	81.98	54.65	65.58	77.48	88.14	81.81	76.05
c8.02	björkelund	■		■				■	71.63	100.00	83.47	65.36	93.23	76.85	64.44	93.51	76.30	68.30	68.30	68.30	78.59	44.24	56.61	75.77	91.56	81.56	69.92
c8.03	xu	■		■				■	71.51	100.00	83.38	65.63	94.07	77.32	65.05	91.23	75.95	66.22	66.22	66.22	78.13	43.77	56.11	73.98	79.71	76.48	69.79
c8.04	stamborg	■		■				■	68.97	100.00	81.63	63.52	88.23	73.86	63.54	88.12	73.84	65.60	65.60	65.60	72.56	42.01	53.21	76.96	83.70	79.89	66.97
c8.05	fernandes	■		■				■	100.00	100.00	100.00	61.64	90.81	73.43	63.55	89.43	74.30	65.10	65.10	65.10	72.78	39.68	51.36	80.21	83.39	81.71	66.36
c8.06	li	■		■				■	63.59	99.95	77.73	55.28	82.28	66.13	55.95	82.99	66.84	57.50	57.50	57.50	66.76	36.06	46.83	70.53	77.67	73.47	59.93
c8.07	zhekova	■		■				■	47.53	100.00	64.43	42.02	79.57	55.00	50.22	80.81	61.94	46.88	46.88	46.88	60.27	27.08	37.37	68.60	63.62	65.58	51.44
c9.00	chen	■			■			■	85.92	100.00	92.42	77.88	92.85	84.71	74.02	91.67	81.91	76.76	76.76	76.76	84.33	59.45	69.74	79.63	91.55	84.45	78.79
c9.01	yuan	■			■			■	82.58	99.80	90.38	75.69	93.06	83.48	71.62	91.39	80.30	74.23	74.23	74.23	82.40	55.60	66.40	77.77	88.30	82.06	76.73
c9.02	björkelund	■			■			■	75.93	100.00	86.32	68.94	93.76	79.46	66.88	93.48	77.97	70.30	70.30	70.30	80.53	47.73	59.93	76.06	91.11	81.66	72.45
c9.03	stamborg	■			■			■	69.08	100.00	81.71	63.52	88.24	73.87	63.56	88.56	74.00	65.89	65.89	65.89	72.93	42.22	53.48	77.43	85.65	80.91	67.12
c9.04	fernandes	■			■			■	100.00	100.00	100.00	61.70	91.45	73.69	63.57	89.76	74.43	65.06	65.06	65.06	72.84	39.49	51.21	80.08	83.21	81.55	66.44
c9.05	li	■			■			■	68.68	100.00	81.43	58.88	81.04	68.20	58.37	80.25	67.59	58.74	58.74	58.74	66.44	38.85	49.03	71.31	76.92	73.72	61.61
c9.06	zhekova	■			■			■	48.82	100.00	65.61	44.12	80.89	57.10	51.79	80.53	63.04	47.84	47.84	47.84	60.37	27.69	37.96	70.49	64.06	66.45	52.70

Table 21: Performance of systems in the *primary* and *supplementary* evaluations for the *closed track* for Chinese.

ID	Participant	Train		Test					MENTION DETECTION			COREFERENCE RESOLUTION															Official
		Syntax		Syntax		Mention Qlty.						MUC			BCUBED			CEAF$_m$			CEAF$_e$			BLANC			
		A	G	A	G	NB	GB	GM	R	P	F	R	P	F$_1$	R	P	F$_2$	R	P	F	R	P	F$_3$	R	P	F	$\frac{F_1+F_2+F_3}{3}$
a0.00	fernandes	■		■		■			62.72	67.00	64.79	43.63	49.69	46.46	62.70	72.19	67.11	55.59	55.59	55.59	52.49	46.09	49.08	63.98	71.91	66.97	54.22
a0.01	björkelund	■		■		■			56.78	64.86	60.55	43.90	52.51	47.82	62.89	75.32	68.54	53.42	53.42	53.42	48.45	40.80	44.30	66.45	74.61	69.63	53.55
a0.02	uryupina	■		■		■			56.47	54.35	55.39	41.33	41.66	41.49	65.77	69.23	67.46	50.82	50.82	50.82	42.43	42.13	42.28	65.58	70.56	67.69	50.41
a0.03	stamborg	■		■		■			56.10	63.28	59.47	39.11	43.49	41.18	61.57	67.95	64.61	50.16	50.16	50.16	44.86	40.36	42.49	66.80	66.94	66.87	49.43
a0.04	chen	■		■		■			56.16	63.95	59.80	38.13	39.96	39.02	60.59	62.51	61.53	47.49	47.49	47.49	41.89	39.84	40.84	66.45	61.84	63.69	47.13
a0.05	zhekova	■		■		■			27.54	80.34	41.02	19.64	62.13	29.85	41.91	90.72	57.33	42.74	42.74	42.74	56.79	24.81	34.53	57.10	79.19	60.65	40.57
a0.06	li	■		■		■			18.17	80.43	29.65	10.77	55.60	18.05	36.17	93.34	52.14	37.03	37.03	37.03	55.45	20.95	30.41	52.91	73.93	54.12	33.53
a1.00	fernandes	■			■	■			65.03	68.71	66.82	46.38	51.78	48.93	63.53	72.37	67.66	56.49	56.49	56.49	52.57	46.88	49.56	64.84	72.97	67.94	55.38
a1.01	björkelund	■			■	■			58.33	64.60	61.30	45.14	52.15	48.39	63.73	74.45	68.68	53.52	53.52	53.52	47.78	41.53	44.44	66.81	73.83	69.65	53.84
a1.12	chen	■			■	■			56.41	63.41	59.70	38.22	39.57	38.89	60.91	62.06	61.48	47.73	47.73	47.73	41.80	40.27	41.02	66.70	61.86	63.78	47.13
a1.13	zhekova	■			■	■			28.00	82.21	41.78	15.47	45.92	23.15	39.22	84.86	53.65	39.52	39.52	39.52	55.10	24.22	33.65	54.13	61.78	55.63	36.82
a2.00	björkelund		■	■		■			61.88	62.52	62.20	46.11	47.66	46.87	65.83	69.74	67.73	53.77	53.77	53.77	45.82	44.33	45.06	67.69	70.71	69.06	53.22
a3.00	björkelund		■		■	■			67.07	62.44	64.67	51.57	49.76	50.65	69.53	69.88	69.71	56.21	56.21	56.21	46.26	47.98	47.11	71.09	72.67	71.85	55.82
a4.00	fernandes	■		■			■		65.34	64.82	65.08	45.18	47.39	46.26	64.56	69.44	66.91	54.88	54.88	54.88	49.73	47.39	48.53	64.28	70.09	66.64	53.90
a4.01	björkelund	■		■			■		57.77	63.74	60.61	44.78	51.47	47.90	63.75	74.27	68.61	53.18	53.18	53.18	47.16	41.24	44.00	66.94	73.43	69.61	53.50
a4.02	stamborg	■		■			■		57.43	64.62	60.81	40.22	44.17	42.10	61.45	67.24	64.22	49.92	49.92	49.92	44.60	40.50	42.46	66.79	66.08	66.42	49.59
a4.03	chen	■		■			■		57.21	62.55	59.76	38.66	39.24	38.95	61.52	61.77	61.65	47.84	47.84	47.84	41.55	40.90	41.22	66.78	61.94	63.87	47.27
a4.04	zhekova	■		■			■		27.48	75.53	40.29	18.75	56.47	28.16	42.67	89.25	57.74	42.57	42.57	42.57	55.53	25.36	34.82	56.61	76.35	59.86	40.24
a4.05	li	■		■			■		52.95	20.71	29.78	20.62	7.78	11.30	79.37	41.21	54.25	33.68	33.68	33.68	21.73	42.87	28.84	54.04	51.10	51.46	31.46
a5.00	fernandes	■			■		■		65.03	68.71	66.82	46.38	51.78	48.93	63.53	72.37	67.66	56.49	56.49	56.49	52.57	46.88	49.56	64.84	72.97	67.94	55.38
a5.01	björkelund	■			■		■		58.29	64.63	61.30	45.14	52.20	48.41	63.71	74.50	68.68	53.52	53.52	53.52	47.80	41.51	44.44	66.81	73.84	69.65	53.84
a5.02	stamborg	■			■		■		57.68	64.18	60.76	40.53	43.98	42.18	61.70	66.75	64.13	49.55	49.55	49.55	44.01	40.47	42.16	65.23	64.89	65.06	49.49
a5.03	chen	■			■		■		56.41	63.45	59.72	38.22	39.59	38.89	60.90	62.07	61.48	47.73	47.73	47.73	41.81	40.26	41.02	66.70	61.86	63.78	47.13
a5.04	zhekova	■			■		■		28.06	82.39	41.87	15.56	46.18	23.28	39.23	84.95	53.67	39.52	39.52	39.52	55.10	24.20	33.63	54.15	61.95	55.66	36.86
a6.00	björkelund		■		■				67.04	62.47	64.67	51.57	49.80	50.67	69.52	69.92	69.72	56.21	56.21	56.21	46.27	47.95	47.10	71.10	72.70	71.86	55.83
a7.00	fernandes	■		■				■	100.00	100.00	100.00	57.25	76.48	65.48	60.27	79.81	68.68	62.56	62.56	62.56	72.61	46.00	56.32	69.03	74.87	71.49	63.49
a7.01	björkelund	■		■				■	61.85	100.00	76.43	49.57	78.62	60.81	55.55	85.35	67.29	59.50	59.50	59.50	70.28	37.99	49.32	70.69	80.85	74.61	59.14
a7.02	zhekova	■		■				■	57.95	100.00	73.38	42.48	80.36	55.58	50.87	89.69	64.92	55.42	55.42	55.42	71.96	34.52	46.66	61.36	82.00	66.12	55.72
a7.03	stamborg	■		■				■	56.13	100.00	71.90	41.99	69.78	52.43	50.45	81.30	62.26	54.00	54.00	54.00	66.16	34.52	45.37	67.37	73.46	69.87	53.35
a7.04	chen	■		■				■	58.29	100.00	73.65	41.72	63.23	50.28	50.00	75.25	60.08	53.16	53.16	53.16	64.60	36.24	46.43	67.15	66.65	66.90	52.26
a7.05	li	■		■				■	35.67	100.00	52.58	22.43	64.62	33.31	38.67	88.07	53.74	42.25	42.25	42.25	60.95	24.36	34.81	55.64	68.52	57.96	40.62
a8.00	fernandes	■			■			■	100.00	100.00	100.00	56.89	76.27	65.17	60.07	80.02	68.62	62.62	62.62	62.62	72.24	45.58	55.90	69.35	75.51	71.93	63.23
a8.01	björkelund	■			■			■	61.05	100.00	75.81	49.17	78.31	60.41	55.51	85.40	67.28	59.41	59.41	59.41	70.01	37.71	49.02	70.97	80.92	74.84	58.90
a8.02	zhekova	■			■			■	65.68	100.00	79.29	45.58	73.27	56.20	52.27	82.35	63.95	55.11	55.11	55.11	70.17	37.54	48.91	59.94	72.07	63.28	56.35
a8.03	stamborg	■			■			■	56.72	100.00	72.38	42.88	70.42	53.30	51.17	80.83	62.67	54.12	54.12	54.12	66.21	34.85	45.66	67.10	72.32	69.29	53.88
a8.04	chen	■			■			■	58.26	100.00	73.63	41.81	63.28	50.36	50.10	75.19	60.13	53.19	53.19	53.19	64.59	36.27	46.46	67.19	66.52	66.85	52.32
a9.00	björkelund		■		■			■	68.50	100.00	81.30	55.21	78.84	64.94	59.85	83.75	69.81	63.12	63.12	63.12	72.24	42.75	53.71	73.35	80.61	76.41	62.82

Table 22: Performance of systems in the *primary* and *supplementary* evaluations for the *closed track* for Arabic.

7.3 Arabic *Closed*

Table 22 shows the performance for the Arabic language in greater detail.

7.3.1 Official Setting

Unlike English and Chinese, none of the system was particularly tuned for Arabic. This gives us an unique opportunity to test the performance variation of a mostly statistical, roughly language independent mechanism. Although, there could possibly be a significant bias that Arabic language brings to the mix. The overall performance for Arabic seems to be about ten points below both English and Chinese. On the mention detection front, most of the systems have a balanced precision and recall, and the drop in performance seems quite steady. *björkelund* has a slight edge on *fernandes* on the MUC, BCUBED and BLANC metrics, but *fernandes* has a much larger lead on both the CEAF metrics, putting it on the top in the official score. We haven't reported the development set numbers here, but another thing to note especially for Arabic is that performance on Arabic test set is significantly better than on the development set as pointed out by *björkelund*. This is probably because of the smaller size of the training set and therefore a higher relative increment over training set. The size of the training set (which is roughly about a third of either Engish or Chinese) also could itself be a factor that explains the lower performance, and that Arabic performance might gain from more data. *chen* did not use development data for the final models. Using that could have increased their score.

7.3.2 Gold Mention Boundaries

The system performance given gold boundaries followed more of the trend in English than Chinese. There was not much improvement over the primary NB evaluation. Interestingly, *chen* uses *gold boundaries* for Chinese so well, but does not get any performance improvement. This might indicate that the technique that helped that system in Chinese does not generalize well across languages.

7.3.3 Gold Mentions

Performance given *gold mentions* seems to be about ten points higher than in the NB case. *björkelund* does well on BLANC metric than *fernandes* even after getting a big hit in recall for mention detection. In absence of *chang*, it seems like *fernandes* is the only one that explicitly adds a constraint for the GM case and gets a perfect mention detection score. All other systems loose significantly on recall.

7.3.4 Gold Test Parses

Finally, providing gold parses during testing does not have much of an impact on the scores.

7.4 All Languages *Open*

Tables 24, 25 and 26, give the performance for the systems that participated in the open track. Not many systems participated in this track, so there is not a lot to observe. One thing to note is that *chen* modified precise constructs sieve to add named entity information in the open track sieve which gave them a point improvement in performance. With *gold mentions* and *gold syntax* during testing the *chen* system performance almost approaches an F-score of 80 (79.79)

7.5 Headword-based and Genre specific scores

Since last year's task showed that there was only some very local difference in ranking between systems scored using the strict boundaries versus the ones using headword based scoring, we did not compute the headword based evaluation.

Owing to space constraints, we cannot present a detailed analysis of the variation across genre. However, since genre variation is important to note, we present the performance of the highest performing system across all the three languages and genres in Table 23. For each language there are three logical performance blocks: i) The *official*, predicted version, with no provided boundaries is the first block; ii) The *supplementary* version with *gold mention boundaries* is the second block; and iii) The third block shows the performance for the *supplementary* version given *gold mentions*.

Looking at the Engish performance on the *official, closed* track, there seems to be a cluster of genre – BC, BN, NW and WB – where the performance is very close to a score of 60. Whereas, genres TC, MZ and PT are increasingly better. Surprisingly, a simplistic look at the individual metrics does indicate a similar trend, except for the CEAF$_e$ score for the TC and WB being somewhat reversed. It so happens that these the two genres — MZ and PT – are professional human translations from a foreign language. As seen earlier, there is not a huge shift in performance when the systems are provided with *gold mention boundaries*. However, when provided with *gold mentions* there is a big improvement in performance across the board. Especially so with MZ genre for which the improvement is more than double (9.5 points) over the improvement in PT genre (3.5 points) with the most notable improvement (of 5 points) in the CEAF$_e$ metric, which also is another indication that this metric does a good job of rewarding correct anaphoric mentions.

Looking at the Chinese performance, we see that the NW genre does particularly worse than all others on the *official, closed* track. The BC genre does somewhat worse than WB, MZ, and TC all of which seem to be around the same ballpark, with BN leading the pack. Again, provided *gold mention bound-*

Genre	Train		Test					MD	COREFERENCE RESOLUTION					Official
	Syntax		Syntax		Mention Qlty.				MUC	BCUBED	CEAF$_m$	CEAF$_e$	BLANC	
	A	G	A	G	NB	GB	GM	F	F$_1$	F$_2$	F	F$_3$	F	$\frac{F_1+F_2+F_3}{3}$
ENGLISH														
Pivot Text [PT]	■		■		■			89.13	82.49	72.66	68.92	54.47	79.20	69.87
Magazine [MZ]	■		■		■			77.70	69.57	77.29	68.88	57.07	81.84	67.98
Telephone Conversation [TC]	■		■		■			79.95	76.75	72.31	62.06	43.22	79.24	64.09
Weblogs and Newsgroups [WB]	■		■		■			78.21	71.66	68.61	59.42	45.24	76.42	61.84
Broadcast News [BN]	■		■		■			74.60	65.15	70.60	60.90	49.52	74.45	61.76
Brodcast Conversation [BC]	■		■		■			75.67	67.54	69.14	57.70	44.99	76.74	60.56
Newswire [NW]	■		■		■			71.24	62.67	71.01	60.61	47.73	75.40	60.47
Pivot Text [PT]	■		■			■		89.50	82.74	72.65	68.98	54.28	79.42	69.89
Magazine [MZ]	■		■			■		77.27	68.68	76.53	67.51	55.63	79.72	66.95
Telephone Conversation [TC]	■		■			■		81.95	78.18	72.53	63.33	44.32	77.99	65.01
Weblogs and Newsgroups [WB]	■		■			■		79.08	72.62	68.94	60.09	45.74	76.46	62.43
Broadcast News [BN]	■		■			■		75.10	65.56	69.98	60.47	49.14	74.10	61.56
Brodcast Conversation [BC]	■		■			■		75.96	67.64	68.51	57.14	44.85	74.81	60.33
Newswire [NW]	■		■			■		70.44	61.63	70.04	59.44	46.57	73.51	59.41
Magazine [MZ]	■		■				■	100.00	82.87	83.10	78.02	66.93	87.00	77.63
Pivot Text [PT]	■		■				■	100.00	86.20	74.30	71.67	59.43	80.12	73.31
Telephone Conversation [TC]	■		■				■	100.00	84.74	75.18	66.29	49.68	77.37	69.87
Weblogs and Newsgroups [WB]	■		■				■	100.00	82.38	71.43	66.08	53.28	77.96	69.03
Newswire [NW]	■		■				■	100.00	74.00	74.41	67.39	53.28	81.03	67.23
Broadcast News [BN]	■		■				■	100.00	74.51	73.31	65.71	52.96	79.15	66.93
Brodcast Conversation [BC]	■		■				■	100.00	77.52	71.49	63.73	50.54	79.59	66.52
CHINESE														
Broadcast News [BN]	■		■		■			78.02	71.71	78.80	68.93	55.87	83.85	68.79
Weblogs and Newsgroups [WB]	■		■		■			79.29	71.30	71.05	60.68	46.81	80.94	63.05
Magazine [MZ]	■		■		■			75.34	70.26	72.32	62.63	46.42	81.34	63.00
Telephone Conversation [TC]	■		■		■			79.79	72.58	71.14	61.16	43.78	76.82	62.50
Brodcast Conversation [BC]	■		■		■			73.80	64.22	67.68	55.38	42.89	72.98	58.26
Newswire [NW]	■		■		■			52.38	49.74	67.97	54.82	43.79	75.63	53.83
Broadcast News [BN]	■		■			■		78.02	71.71	78.80	68.93	55.87	83.85	68.79
Weblogs and Newsgroups [WB]	■		■			■		79.29	71.30	71.05	60.68	46.81	80.94	63.05
Magazine [MZ]	■		■			■		75.34	70.26	72.32	62.63	46.42	81.34	63.00
Telephone Conversation [TC]	■		■			■		79.79	72.58	71.14	61.16	43.78	76.82	62.50
Brodcast Conversation [BC]	■		■			■		73.80	64.22	67.68	55.38	42.89	72.98	58.26
Newswire [NW]	■		■			■		52.38	49.74	67.97	54.82	43.79	75.63	53.83
Broadcast News [BN]	■		■				■	100.00	81.03	81.34	75.07	62.99	86.18	75.12
Telephone Conversation [TC]	■		■				■	100.00	87.31	77.80	71.01	59.44	78.78	74.85
Weblogs and Newsgroups [WB]	■		■				■	100.00	80.36	72.46	64.49	51.93	81.10	68.25
Magazine [MZ]	■		■				■	100.00	75.18	73.12	65.90	48.81	84.17	65.70
Brodcast Conversation [BC]	■		■				■	100.00	76.42	70.01	61.75	49.81	74.14	65.41
Newswire [NW]	■		■				■	100.00	51.42	67.83	55.29	43.81	76.53	54.35
ARABIC														
Newswire [NW]	■		■		■			64.79	46.46	67.11	55.59	49.08	66.97	54.22
Newswire [NW]	■		■			■		65.08	46.26	66.91	54.88	48.53	66.64	53.90
Newswire [NW]	■		■				■	100.00	65.48	68.68	62.56	56.32	71.49	63.49

Table 23: Per genre performance for *fernandes* on the *closed*, *official* and *supplementary* evaluations.

aries there is very little or no change in performance. And, when given the *gold mentions* the performance again shoots up by a significant margin. Here again, we see that the delta improvement in one particular genre TC – is much higher (12 points) than in BN (6 points), and once again the most improvement among all the metrics happens to be for the CEAF$_e$. Extremely surprising is the fact that the NW genre shows the lowest improvement among all genre. In fact, the performance drops for the BCUBED metric. This might have something to do with the fact that Chinese NW genre gets the lowest ITA among all other (see Table 1), but then the better scoring TC genre which has the second lowest ITA does considerably better (leading by roughly 10 points in the *official* setting, and 20 points in the *gold mentions* settings with respect to the TC genre). It could also be possible that this has something to do with the fact (and pointed out earlier when discussing *chen*'s results) that there is some overlapping mentions that were mistakenly included in the release.

As for Arabic, since there was only one NW genre, there is nothing more to be analyzed. We plan to report more detailed tables and analysis on the task webpage.

8 Comparison with CoNLL-2011

Table 27 shows the performance of the systems on CoNLL-2011 test subset which included only the English portion of OntoNotes v4.0. For the English subset, the size of training data in CoNLL-2011 was roughly 76% of CoNLL-2012 training data (1M vs 1.3M words respectively). Although the models used to generate this table were trained on the CoNLL-2012 English data and therefore on about 200K more words, it is still a small fraction of the total training data. In the past, coreference scores have shown to asymptote after a small fraction of the total training data. Therefore, the 5% absolute gap between the best performing systems of last year can be attributed to algorithmic improvement, and possibly better rules. Given that a 200K data addition to a 1M word corpus is unlikely to help identify novel rules, and given that *björkelund* reported adding (about 160K) development data (to the training portion) to train the final model had very little improvement in performance over using just the training data by itself, the possibility that the gain is from algorithmic improvements seems even more plausible.

It is interesting to note that although the winning

Participant	Train		Test					MENTION DETECTION			COREFERENCE RESOLUTION															Official
	Syntax		Syntax		Mention Qlty.						MUC			BCUBED			CEAF$_m$			CEAF$_e$			BLANC			
	A	G	A	G	NB	GB	GM	R	P	F	R	P	F$_1$	R	P	F$_2$	R	P	F	R	P	F$_3$	R	P	F	$\frac{F_1+F_2+F_3}{3}$
xiong	■		■		■			75.22	72.23	73.69	64.08	63.57	63.82	66.47	70.69	68.52	57.23	57.23	57.23	45.09	45.64	45.36	71.12	77.90	73.94	59.23
xiong	■			■	■			77.85	73.44	75.58	67.03	65.27	66.14	68.03	71.14	69.55	58.58	58.58	58.58	45.60	47.53	46.54	72.03	78.58	74.78	60.74

Table 24: Performance of systems in the *primary* and *supplementary* evaluations for the *open* track for English.

Participant	Train		Test					MENTION DETECTION			COREFERENCE RESOLUTION															Official
	Syntax		Syntax		Mention Qlty.						MUC			BCUBED			CEAF$_m$			CEAF$_e$			BLANC			
	A	G	A	G	NB	GB	GM	R	P	F	R	P	F$_1$	R	P	F$_2$	R	P	F	R	P	F$_3$	R	P	F	$\frac{F_1+F_2+F_3}{3}$
chen	■		■		■			71.45	73.45	72.44	62.48	67.08	64.70	71.21	78.35	74.61	63.48	63.48	63.48	53.64	49.10	51.27	75.15	84.29	78.94	63.53
yuan	■		■		■			73.71	63.97	68.49	63.67	58.48	60.96	74.04	72.16	73.09	60.05	60.05	60.05	46.75	51.52	49.02	74.32	77.99	76.02	61.02
xiong	■		■		■			39.47	67.55	49.82	30.00	51.20	37.83	49.37	77.45	60.30	42.71	42.71	42.71	46.10	28.12	34.93	57.98	67.08	60.59	44.35
chen	■			■	■			83.50	80.44	81.95	74.77	74.93	74.85	77.14	80.80	78.93	70.13	70.13	70.13	58.64	58.46	58.55	78.87	86.63	82.24	70.78
yuan	■			■	■			83.92	69.85	76.24	74.24	65.11	69.37	78.61	73.74	76.10	64.84	64.84	64.84	49.41	58.70	53.66	77.60	79.45	78.49	66.38
xiong	■			■	■			42.78	71.10	53.42	32.57	53.89	40.60	49.50	77.38	60.37	43.66	43.66	43.66	47.04	28.83	35.75	58.24	67.37	60.90	45.57
chen	■		■			■		82.39	80.11	81.24	73.50	74.28	73.88	76.30	80.49	78.34	69.40	69.40	69.40	58.22	57.32	57.77	78.44	86.39	81.88	70.00
chen	■			■		■		83.50	80.44	81.95	74.77	74.93	74.85	77.14	80.80	78.93	70.13	70.13	70.13	58.64	58.46	58.55	78.87	86.63	82.24	70.78
chen	■		■				■	84.80	100.00	91.77	78.12	93.19	84.99	75.04	91.59	82.50	77.50	77.50	77.50	84.03	59.17	69.44	81.46	90.73	85.41	78.98
chen	■			■			■	85.71	100.00	92.31	79.07	93.59	85.72	75.83	91.94	83.11	78.26	78.26	78.26	84.77	60.42	70.55	81.71	91.00	85.67	79.79

Table 25: Performance of systems in the *primary* and *supplementary* evaluations for the *open* track for Chinese.

Participant	Train		Test					MENTION DETECTION			COREFERENCE RESOLUTION															Official
	Syntax		Syntax		Mention Qlty.						MUC			BCUBED			CEAF$_m$			CEAF$_e$			BLANC			
	A	G	A	G	NB	GB	GM	R	P	F	R	P	F$_1$	R	P	F$_2$	R	P	F	R	P	F$_3$	R	P	F	$\frac{F_1+F_2+F_3}{3}$
xiong	■		■		■			55.08	52.75	53.89	28.20	28.43	28.31	60.89	62.81	61.83	47.31	47.31	47.31	43.12	42.82	42.97	57.05	60.75	58.46	44.37
xiong	■			■	■			57.55	52.98	55.17	30.99	30.10	30.54	62.16	62.55	62.36	47.73	47.73	47.73	42.48	43.59	43.03	57.78	61.39	59.20	45.31

Table 26: Performance of systems in the *primary* and *supplementary* evaluations for the *open* track for Arabic.

Participant	Train		Test					MENTION DETECTION			COREFERENCE RESOLUTION															Official
	Syntax		Syntax		Mention Qlty.						MUC			BCUBED			CEAF$_m$			CEAF$_e$			BLANC			
	A	G	A	G	NB	GB	GM	R	P	F	R	P	F$_1$	R	P	F$_2$	R	P	F	R	P	F$_3$	R	P	F	$\frac{F_1+F_2+F_3}{3}$
2011 best	■		■		■			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
fernandes	■		■		■			70.12	81.27	75.28	62.74	73.46	67.68	65.03	78.05	70.95	60.61	60.61	60.61	54.18	42.50	47.63	75.31	78.53	76.80	62.09
martschat	■		■		■			71.96	73.57	72.76	62.80	66.23	64.47	67.09	74.50	70.60	58.84	58.84	58.84	48.05	44.55	46.23	73.17	78.04	75.32	60.43
björkelund	■		■		■			70.99	74.91	72.90	62.25	67.72	64.87	65.66	75.32	70.16	57.99	57.99	57.99	48.12	42.67	45.23	72.81	77.64	74.94	60.09
chang	■		■		■			69.94	73.36	71.61	62.51	65.54	63.99	67.76	71.97	69.80	57.49	57.49	57.49	45.86	42.82	44.29	75.35	74.13	74.72	59.36
chen	■		■		■			73.13	69.56	71.30	60.84	60.70	60.77	67.34	71.37	69.30	57.47	57.47	57.47	45.98	46.13	46.05	70.89	78.69	74.06	58.71
stamborg	■		■		■			72.83	69.78	71.27	63.47	61.16	62.29	69.10	69.19	69.14	55.37	55.37	55.37	41.89	44.13	42.98	72.86	75.67	74.16	58.14
chunyang	■		■		■			73.14	69.31	71.17	61.78	60.57	61.17	67.43	70.31	68.84	56.62	56.62	56.62	44.48	45.70	45.08	71.31	76.91	73.72	58.36
yuan	■		■		■			70.44	68.87	69.64	58.99	60.09	59.53	66.66	71.38	68.94	56.95	56.95	56.95	45.48	44.38	44.92	72.20	78.69	74.95	57.80
shou	■		■		■			73.26	69.16	71.15	61.02	59.24	60.12	66.14	68.34	67.22	54.88	54.88	54.88	43.52	45.35	44.42	69.00	73.39	70.92	57.25
xu	■		■		■			58.90	82.61	68.77	55.14	73.28	62.93	60.50	75.93	67.35	52.81	52.81	52.81	48.71	32.17	38.75	72.80	70.05	71.30	56.34
uryupina	■		■		■			69.64	66.80	68.19	58.92	57.72	58.31	65.05	68.26	66.62	52.25	52.25	52.25	40.71	41.83	41.26	68.07	72.25	69.89	55.40
yang	■		■		■			61.51	70.97	65.90	52.15	61.88	56.60	60.37	73.02	66.09	51.07	51.07	51.07	44.69	35.98	39.87	66.20	71.27	68.29	54.19
xinxin	■		■		■			71.16	50.98	59.40	52.88	39.37	45.13	68.73	56.16	61.81	44.68	44.68	44.68	31.27	43.50	36.38	65.71	64.79	65.23	47.77
zhekova	■		■		■			63.16	65.73	64.42	50.64	49.51	50.07	61.71	56.53	59.01	43.40	43.40	43.40	34.01	35.09	34.54	65.49	58.37	60.21	47.87
li	■		■		■			43.39	84.81	57.40	36.45	70.53	48.06	43.11	81.37	56.36	41.88	41.88	41.88	49.52	22.69	31.12	62.31	67.02	64.12	45.18

Table 27: Performance of all the systems on the CoNLL-2011 portion (English) of the CoNLL-2012 test set.

system in the CoNLL-2011 task was a completely rule-based one, modified version of the same system used by *shou* and *xiong* ranked close to 10. This does indicate that a hybrid approach has some advantage over a purely rule-based system. Improvement seems to be mostly owing to higher precision in mention detection, MUC, BCUBED, and higher recall in CEAF$_e$

9 Conclusions

In this paper we described the anaphoric coreference information and other layers of annotation in the OntoNotes corpus, over three languages — English, Chinese and Arabic — and presented the results from an evaluation on learning such unrestricted entities and events in text. The following represents our conclusions on reviewing the results:

- Most top performing systems used a hybrid-approach combining rule-based strategies with machine learning. Rule-based approach does seem to bring a system to a close-to-best performance region. The most significant advantage of the rule-based approach seems to be that it captures most confident links before considering less confident ones. Discourse information when present is quite helpful to disambiguate pronominal mentions. Using information from appositives and copular constructions seems beneficial to bridge across various lexicalized mentions. It is not clear how much more can be gained using further strategies. 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.
- Most top performing systems did significant feature engineering – expecially a heavy use of lexicalized features, which was possible given the size of the corpus, and performed feature selection.
- It might be possible that the Chinese accuracy with gold boundaries and mentions is better because the distribution of mentions across the various genres is different, and if there are more mentions in better scoring genres, then the performance would improve overall.
- Gold parse during testing does seem to help quite a bit. Gold boundaries are not of much significance for English and Arabic, but seem to be very useful for Chinese. The reason probably has some roots in the parser performance gap for Chinese.
- 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, and has to be done carefully.
- 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 the CEAF$_e$ 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.

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 v5.0 corpus freely and timely available in training/development/test sets to the participants. Emili Sapena, who graciously allowed the use of his scorer implementation. Hwee Tou Ng and his student Zhi Zhong for training the word sense models and providing outputs for the training/development and test sets. Slav Petrov and Dan Klein for letting us use their parser. Additionally, we are indebted to Slav for his help in retraining the parser for Arabic. Alessandro Moschitti and Olga Uryupina have been partially funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under the grant number 288024 (LIMOSINE).

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.

Elizabeth Baran and Nianwen Xue. 2011. Singular or Plural? Exploiting Parallel Corpora for Chinese Number Prediction. In *Proceedings of Machine Translation Summit XIII*.

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.

Jie Cai, Eva Mujdricza-Maydt, and Michael Strube. 2011a. Unrestricted coreference resolution via global hypergraph partitioning. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 56–60, Portland, Oregon, USA, June. Association for Computational Linguistics.

Shu Cai, David Chiang, and Yoav Goldberg. 2011b. Language-independent parsing with empty elements. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 212–216, Portland, Oregon, USA, June. Association for Computational Linguistics.

Wendy W Chapman, Prakash M Nadkarni, Lynette Hirschman, Leonard W D'Avolio, Guergana K Savova, and Ozlem Uzuner. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. *Journal of American Medical Informatics Association*, 18(5), September.

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.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *Proceedings of LREC*.

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

Ryan Gabbard. 2010. *Null Element Restoration*. Ph.D. thesis, University of Pennsylvania.

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*.

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

Lynette Hirschman, Patricia Robinson, John Burger, and Marc Vilain. 1998. Automating coreference: The role of annotated training data. In *Proceedings of AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*.

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.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34, Portland, Oregon, USA, June. Association for Computational Linguistics.

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.

Mohamed Maamouri and Ann Bies. 2004. Developing an arabic treebank: Methods, guidelines, procedures, and tools. In Ali Farghaly and Karine Megerdoomian, editors, *COLING 2004 Computational Approaches to Arabic Script-based Languages*, pages 2–9, Geneva, Switzerland, August 28th. COLING.

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.

Eugene W. Myers. 1986. An O(ND) difference algorithm and its variations. *Algorithmica*, 1(2):251—-266.

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.

David S. Pallett. 2002. The role of the National Institute of Standards and Technology in DARPA's Broadcast News continuous speech recognition research program. *Speech Communication*, 37(1-2), May.

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.

Martha Palmer, Olga Babko-Malaya, Ann Bies, Mona Diab, Mohammed Maamouri, Aous Mansouri, and Wajdi Zaghouani. 2008. A pilot arabic propbank. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 28-30.

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

Slav Petrov and Dan Klein. 2007. Improved Inferencing for Unlexicalized Parsing. In *Proc of HLT-NAACL*.

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.

Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nate Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. 2010. A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501, Cambridge, MA, October. Association for Computational Linguistics.

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.

William 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.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong 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.

Ozlem Uzuner, Andreea Bodnari, Shuying Shen, Tyler Forbush, John Pestian, and Brett R South. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of American Medical Informatics Association*, 19(5), September.

Yannick Versley, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. BART: A modular toolkit for coreference resolution. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, May.

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)*.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette 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, Mitchell Marcus, Martha Palmer, 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: DARPA Global Autonomous Language Exploitation*. Springer.

Nianwen Xue and Martha Palmer. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15(1):143–172.

Nianwen Xue, Fei Xia, Fu dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207–238.

Nianwen Xue. 2008. Labeling Chinese Predicates with Semantic Roles. *Computational Linguistics*, 34(2):225–255.

Yaqin Yang and Nianwen Xue. 2010. Chasing the ghost: recovering empty categories in the chinese treebank. In *Proceedings of Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, Beijing, China.

Wajdi Zaghouani, Mona Diab, Aous Mansouri, Sameer Pradhan, and Martha Palmer. 2010. The revised arabic propbank. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 222–226, Uppsala, Sweden, July.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 System Demonstrations*, pages 78–83, Uppsala, Sweden.