# Machine Learning for Mention Head Detection in Multilingual Coreference Resolution

**Desislava Zhekova**
CIS, University of Munich
zhekova@cis.uni-muenchen.de

**Sandra Kübler**
Indiana University
skuebler@indiana.edu

## Abstract

This work introduces a machine learning approach to the identification of mention heads needed for multilingual coreference resolution (MCR). We evaluate the method and compare it to a heuristic baseline and a rule-based approach, which are widely used in coreference resolution systems. We use the CoNLL-2012 shared task data sets, which include data for Arabic, Chinese, and English. We show that for MCR, machine learning offers a competitive, flexible, and robust solution for mention head detection.

## 1 Introduction

Coreference Resolution (CR) aims to detect all linguistic expressions in a given discourse that refer to real world entities. Such expressions are generally called *mentions*. They need to be grouped into equivalence classes so that each class contains only mentions that refer to the same entity. The classes are called *coreference chains*. The task of CR includes not only the identification of coreference links between mentions, but also the detection of the mentions themselves. This subtask of CR has not been a main topic of interest, since most of the standard data sets for CR contained gold mention information. This situation changed in the most recent shared tasks on the topic of CR: SemEval-2010 Task 1 (Recasens et al., 2010), CoNLL-2011 (Pradhan et al., 2011) and CoNLL-2012 (Pradhan et al., 2012). The data distributed by these tasks included syntactic annotations, and it was considered an integral part of the task for the participating systems to develop their own methods to detect mention boundaries.

Statistical approaches to the CR problem often recast the task to a binary classification exercise. For the latter, coreference is represented by a decision model, such as the mention-pair model (Soon et al., 2001). The mention-pair model, which is the most widely used model for CR, pairs the anaphor with a potential antecedent, and determines whether they are coreferent or not. Since the decisions are taken independently for each possible antecedent, a global heuristic can be used to decide between multiple positive decisions or in cases where no antecedent was found.

The use of the mention-pair model implies that an instance consists of a pair of mentions, and, since vectors for machine learning (ML) need to be of a fixed length, each mention is generally represented by its syntactic head, plus informative features that describe the phrases and their context. As a consequence, there is an additional subtask of CR that needs to be performed before the actual resolution process: *mention head detection* (MHD). This is usually done by the use of simple heuristics or manually defined sets of rules (see section 2). In this work, we will investigate a novel ML method for multilingual MHD.

Multilinguality has presented additional issues to the coreference task, which were discussed and addressed by the two multilingual shared tasks on the topic SemEval-2010 Task 1 and CoNLL-2012. In general, MCR is faced with the same problems as monolingual CR: we have to optimize the 3 main stages in CR, the actual detection of mentions (MD), the detection of the syntactic heads of the latter and classification, based on a selection of features that can represent the phenomena. In our current work, we assume a mention-pair coreference model.

Identifying the head of a phrase is closely related to detecting the grammatical structure of sentences. Thus, the annotation layers provided in the two shared tasks led to the development of successful methods for MD that were mostly based on the underlying syntactic structure of the sentences. To our knowledge most state-of-the-art CR systems have not regarded MHD as a stand-

*Proceedings of Recent Advances in Natural Language Processing*, pages 747–754,
Hissar, Bulgaria, 7-13 September 2013.

alone subtask of CR, but rather as part of the feature extraction process. Since mentions often correspond to NPs, most approaches use variants of head finding rules, which were made popular by Collins (1999). Such rules are manually written and specify where to find the head for an individual syntactic category.

In this work, we pursue the goal of MCR in the sense that we are developing an architecture that allows CR for multiple languages with only a minimal adaptation to the individual language. This means that we also need a multilingual approach to MHD that does not require the development of head-finding rules for every language to be added to the system. Thus, we introduce a novel method for MHD based on a ML approach, and we compare it to two widely used approaches.

In section 2, we give a short overview of the state of the art, then we present the problems with respect to multilinguality and the head detection problem (section 3). In section 4, we describe the two existing approaches to MHD and propose our own ML method. Section 5 describes the data set and evaluation settings and presents a comparison of the ML approach with respect to the other two approaches. In section 6 we conclude our observations and delineate future directions for this task.

## 2 Related Work

While there is a bulk of literature on CR for English (Soon et al., 2001; Ng and Cardie, 2002; Ng, 2007, for example), MCR has only been addressed recently. The majority of work in this area was carried out in the context of the two shared tasks, the SemEval-2010 (Recasens et al., 2010) and the CoNLL-2012 (Pradhan et al., 2012) tasks. We focus on MHD for the data from CoNLL-2012.

The majority of the systems participating in the two shared tasks used approaches that were fairly language dependent with respect to MHD. In the context of the CoNLL-2012 task, the systems by Chen and Ng (2012), Martschat et al. (2012), and Uryupina et al. (2012) used manually created sets of rules, based on head-finding models following (Collins, 1999). This means that every language other than English, which is targeted by these systems, would need other, language specific, sets of rules. Björkelund and Farkas (2012) employed Choi and Palmer (2010)'s percolation rules for Arabic and English and the rules of Zhang and Clark (2011) for Chinese. Li et al. (2012) used the head-finding rules from Penn2Malt, for English and for Chinese. The system by Martschat et al. (2012) relies on the Stanford SemanticHeadFinder (also an implementation of the rules by Collins (1999)) for English while the head detection for Chinese is provided by the SunJurafskyChineseHeadFinder (an implementation of the rules presented by Sun and Jurafsky (2004)). Martschat et al. (2012) did not work on CR for Arabic.

Uryupina et al. (2012) created their own heuristic rules for the Arabic and Chinese; for English, they used Collins (1999)'s rules. For Arabic, the first noun/pronoun was selected as head; in Chinese, the last noun/pronoun was chosen as the head. Uryupina et al. (2012) also made the observation that the absence of expert linguistic knowledge can become an important obstacle when rules are to be developed manually for each separate language. Additionally, depending on the language, the collection of such rules may be a rather expensive task.

## 3 Issues in Multilingual MHD

The concept of a mention is closely related to NPs in syntax. The reason for this relation is that CR at present focuses on entities and often ignores event coreference. As a consequence, finding the head of a mention generally corresponds to identifying the syntactic head of the corresponding NP. The major difference lies in the fact that mentions often correspond to maximal rather than to base NPs.

If we approach the task of finding the mention heads by identifying syntactic heads, the task would be trivial if we had a full syntactic analysis, as provided in X-bar theory (Chomsky, 1970; Jackendoff, 1977) or in *head-driven phrase structure grammar* (Sag et al., 2003; Levine and Meurers, 2006). However, treebanks are generally annotated in a more surface-oriented and flat annotation, in which heads of phrases are often not marked as such. The Penn Treebank (Marcus et al., 1993), which is the standard for training statistical parsers for English, for example, uses a flat annotation scheme for NPs, as shown in the examples in (1). The annotation in the Penn Treebank for English also served as the model for the annotations in the Penn Arabic and Chinese treebanks.

(1) a. [$_{\text{NP}}$ The average seven-day compound yield]
b. [$_{\text{NP}}$ [$_{\text{NP}}$ the ceiling] [$_{\text{PP}}$ on [$_{\text{NP}}$ government debt]]]
c. [$_{\text{NP}}$ [$_{\text{NP}}$ executives] and [$_{\text{NP}}$ their wives]]

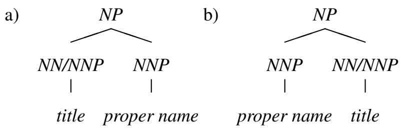

Figure 1: The structure of NPs with titles; with the head in phrase-initial or -final position.

Phrase directionality, which describes the position of the syntactic head in a phase, is fairly regular for most languages, which is mainly why MHD is generally performed via heuristics or language dependent sets of rules. The languages in the CoNLL-2012 shared task represent a good variation of directionalities: Arabic is a consistently head-initial language; Chinese is a consistently head-final language; and English represents a language with mixed directionality since it places specifiers before the head and heavier constituents, such as prepositional phrases or relative clauses, after the syntactic head. Thus, English is the most difficult case: it requires knowledge of the internal structure of the NP in order to correctly identify the head of a higher-order NP, which is non-trivial to capture in a heuristic or in rules.

In the context of the CoNLL-2012 shared task, one simple type of NP that is difficult to capture by heuristics across languages consists of phrases containing a combination of titles, such as *Mr.*, or *Dr.*, and proper names. In all three data sets, proper names are part-of-speech (POS) tagged as NNP, titles can be tagged as either NN or NNP depending on the language: In English, titles are NNP, in Chinese NN, and in Arabic, they are NOUN_PROP in the gold annotations, but the automatically assigned tag is NN. Generally, there are two possibilities where to place titles: either directly before or directly after the proper name, which is visually represented in figure 1. In both cases, the proper name is the head of the full NP. While in Arabic and English, titles are placed before the proper names, in Chinese, they are in phrase-final position. A simple heuristic approach to MHD, using either the first or the last token in the mention, would not capture the proper token as a head of such phrases. For Arabic, for example, as a head-initial language, the heuristic will pick the first token of the phrase to be the head. However, in that position, Arabic places the titles and not the proper names. In contrast, for Chinese, the last token will be selected, but this language places the titles after the names.

Titles and proper names are not the only phrase type that is difficult to be covered by heuristics. Other such types include full person names with the use of given and surname or more complex cases, involving coordinated phrases that need to elicit more than one head. Such complex cases cannot be covered by a simple heuristic, but rather need to be defined via language dependent rules in order to be captured properly across languages. However, as mentioned before, this requires linguistic knowledge of the language in question.

## 4 Methods for Multilingual MHD

In this section, we discuss 2 baseline methods and our novel ML method.

**Heuristic MHD (HeuristicH)** Detecting the head of the phrase via a heuristic considers only the predominant language directionality. For example, since Arabic is consistently head-initial, the heuristic will choose the first noun/pronoun to be the head of each NP. For head-final languages, the last noun/pronoun is selected. Since English has a mixed directionality, we treat it as a head-initial language. We are aware that this is not a good fit for English, but we aim at modeling lesser resourced languages with mixed directionality, for which no language specific knowledge is available. We employ the heuristic without improvement as a language independent baseline for which the only knowledge needed is the predominant directionality of the NPs in that language.

**Rule-based MHD (RuleH)** The rule-based approach uses a set of rules: every set consists of language dependent rules that cover MHD for coordinated phrases and the occurrence of proper names and titles. For English, we include a rule defining the head to be the last noun/pronoun in a sequence of nouns/pronouns, which addresses the problem of nominal premodification. We also restrict the search for the head to words before postmodifier clauses. Our rule set is similar to the one by Collins (1999). However, since we extract heads for mentions rather than for (often nested) phrases, we modified the rules so that they consider context to account for the mixed directionality of English NPs (i.e., the search stops at e.g. prepositions).

**Machine Learning for MHD (MLH)** Our machine learning method is based on memory-based learning (MBL), which has been shown to have a

| # | Feature Description |
|---|---|
| 1 | the target token |
| 2 | part-of-speech tag of the target token |
| 3 | part-of-speech tag of token$_{-1}$ |
| 4 | part-of-speech tag of token$_{+1}$ |
| 5 | Y if it is the only token in the mention; else N |
| 6 | Y if it is not in a PP, SBAR, VP, S; else N |
| 7 | Y if it is the first token in the mention; else N |
| 8 | Y if it is the last token in the mention; else N |
| 9 | Y if the target token is a noun |
| 10 | Y if the target token is a pronoun |
| 11 | Y if the target token is a noun or a pronoun |
| 12 | Y if the target token is followed by a noun |
| 13 | Y if the target token is followed by a pronoun |
| 14 | Y if the following token is possessive and the last token in the mention |

Table 1: The 14 features for the MLH classifier.

good bias for a variation of NLP problems (Daelemans and van den Bosch, 2005), more specifically TiMBL (Daelemans et al., 2010), an efficient implementation of the $k$-nearest neighbor ($k$-NN) approach. MBL classifies a new instance based on the $k$ closest examples from the training set. If the $k$ nearest examples are distributed over different classes, the majority of the set is used. We do not perform parameter optimization.

In the current task of MHD, we create an instance for every word in a mention, and decide for this word whether it is the head of the mention or not. As mentions we select the set of *gold* mentions provided by the task. Since mention head information is not provided in standard data distributions and was not included in the CoNLL-2012 data, we manually annotated a small data set.

In order to create the training/test data sets, all mentions from the training data are extracted, and each of the tokens for each of the mentions is represented as a feature vector containing information about the context of the given token in the current mention. As features, we collect 14 language independent values, listed in table 1. The features are extracted from the POS annotation layer.

One problem that is not handled by the MLH approach is that the tokens are classified individually, i.e., it is possible that more than one token is classified as the head. However, mentions that do not contain coordinating conjunctions should be assigned exactly one head. Correspondingly, the existence or type of the coordinating conjunction could be used in order to restrict the output of the classifier, which can be also regulated via a weighted classification procedure. In our work, we did not postprocess the output of the classifier, i.e., the output may contain multiple heads per mention.

## 5 Mention Head Detection Experiments

The evaluation of MHD is not a trivial task, since as noted before, mention heads are not included in standard linguistic annotation layers. It is also not part of the evaluation software provided by the shared tasks.

First, in section 5.1, we describe the data set and the experimental setup, including the CR system that we use. Then, we perform two different types of evaluation: In section 5.2, we assess the performance of the three MHD methods on the manually annotated data sets in an intrinsic evaluation, without integrating them into the full CR pipeline. And in section 5.3, we perform an extrinsic evaluation by using each of the three methods in an MCR system and compare the CR performance achieved by the approaches.

### 5.1 Data Set and Experimental Setup

For the following experiments, we used the CoNLL-2012 training and test data sets. In order to be able to assemble training data for the ML approach, we manually annotated a subset of the data for each of the three languages in the task. The data for Arabic includes an excerpt of 42 documents for training and 8 for testing. For English, we consider 100 documents for training and 20 documents for testing. Finally, for Chinese, 84 documents are annotated as a training set, and 16 are used as a test set. Note that Arabic has a significantly lower number of annotated documents, which is not only the result of its smaller data sets, but rather a consequence of the fact that Arabic has a highly NP-rich syntactic structure, which accounts for substantially more training instances per document than for English and Chinese. The annotations for English were performed by the first author, the ones for Chinese and Arabic by linguistically educated native speakers. The mentions used for the experiments are *gold* mentions, thus only coreferent mentions. Overall, the number of instances extracted are similar across all three languages. On average, the annotation of the data set required approximately two person-days per language.

For the intrinsic evaluation in section 5.2, we calculate precision, recall, and F$_1$-score. For the extrinsic evaluation in section 5.3, we asess the results in the full pipeline of a MCR system. We use the UBIU architecture (Zhekova and Kübler, 2010). UBIU is based on the mention-pair model

| language | metric | HeuristicH | RuleH | MLH |
|---|---|---|---|---|
| AR | R | 0.79 | 0.83 | **0.85** |
| excerpt | P | 0.87 | 0.88 | **0.91** |
| | $F_1$ | 0.83 | 0.85 | **0.88** |
| EN | R | 0.65 | **0.92** | 0.87 |
| excerpt | P | 0.70 | 0.97 | **0.98** |
| | $F_1$ | 0.67 | **0.95** | 0.92 |
| ZH | R | 0.84 | 0.96 | **0.97** |
| excerpt | P | 0.98 | 0.98 | **0.99** |
| | $F_1$ | 0.90 | 0.97 | **0.98** |

Table 2: MHD for excerpt data for all languages, Arabic (AR), English (EN), and Chinese (ZH), for all spans of mentions.

and uses TiMBL for classification. Since we are more interested in the effects of the MHD methods on the full CR system rather than in the optimal performance that can be achieved by UBIU, we do not aim at language dependent system optimization on any system component. We use the official CoNLL-2012 scorer, which provides five evaluation metrics: MUC (Vilain et al., 1995), $B^3$ (Bagga and Baldwin, 1998), the two variants of CEAF (Luo, 2005), CEAF$_E$ and CEAF$_M$, and BLANC (Recasens and Hovy, 2011). For comparison, we calculate a TOTAL score as the average of the F-score of all metrics.

### 5.2 Intrinsic Evaluation

The results in table 2 show an interesting outcome: HeuristicH, which requires only minimal language specific knowledge, leads to the lowest performance across all three languages, with an F-score of 0.83 for Arabic, 0.90 for Chinese, and 0.67 for English. This outcome shows that for Arabic and Chinese, we reach a very competitive performance with a rather simple heuristic. Remember that both Arabic and Chinese have a clearly unidirectional NP structure. For English, however, with its mixed directionality in NPs, the results are far below the results for the other two languages, with a difference of 23 percent points between Chinese and English. This difference is a direct consequence of the various issues specific to English that we introduced in section 3, such as nominal premodification. Therefore, HeuristicH should only be used when it is known that a language is unidirectional. Even in such cases, we cannot expect a high performance in every case.

The rule-based approach partially addresses the shortcomings of the HeuristicH baseline. It achieves an F-score of 0.85 for Arabic, 0.95 for English, and 0.97 for Chinese. This shows that we can reach very reliable results for English and Chinese; especially for English, which shows an improvement by 28 percent points, from an F-score of 0.67 to 0.95. For Arabic, however, the gain from the heuristic to the rule-based approach is minimal: it only gains 2 percent points, and it is far from reaching 90%.

The results for MLH show that this method is highly competitive: For Arabic (with an F-score of 0.88) and Chinese (with an F-score of 0.98), the ML approach reaches the best performance on the task. For English, the overall performance is 0.92, which is only 3 percent points lower than for the rule-based variant. Moreover, the scores for this language show that RuleH reaches a higher recall while precision is better for MLH. Part of the low recall for English may be due to the fact that the training set is restricted in size, which is detrimental for English since there the task is more difficult because of the mixed directionality in NPs.

Note also that overall, for all languages, precision is always higher than recall, which allows the conclusion that our simplistic approach in the ML method, allowing more than one head, does not harm the method's performance. Overall, we can conclude that the MLH approach is capable of learning the different directionalities, and it is highly competitive, especially given that it is a language independent method that can be employed for any language for which POS information is provided, given a small annotated data set.

### 5.3 Extrinsic Evaluation

For the extrinsic evaluation, we integrate all methods for MHD into the complete MCR pipeline. This shows whether the MHD methods have an effect on CR. The results are listed in table 3. As upper bound, we use *gold standard* heads. The results show the same trends as in our intrinsic evaluation: HeutisticH consistently reaches the lowest scores across all languages, with TOTAL scores as follows: Arabic: 30.54, English: 40.10 and Chinese: 37.53.

RuleH again achieves higher scores in comparison to the heuristic across all languages. This again confirms our observations that HeuristicH is not a good fit for a multilingual environment. RuleH reaches a TOTAL score of 31.74 for Arabic, 48.40 for English and 48.21 for Chinese. This leads altogether to the best observed performance for the English language. However, for Arabic and Chinese, MLH once more performs best with

| | | AR | | | EN | | | ZH | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | R | P | $F_1$ | R | P | $F_1$ | R | P | $F_1$ |
| HeuristicH | MD | 4.85 | 43.05 | 8.72 | 42.18 | 51.39 | 46.33 | 58.73 | 40.43 | 47.89 |
| | MUC | 1.70 | 18.18 | 3.11 | 24.31 | 28.87 | 26.39 | 42.29 | 30.33 | 35.32 |
| | $B^3$ | 31.82 | 94.60 | 47.63 | 52.29 | 62.17 | 56.81 | 61.54 | 45.82 | 52.53 |
| | $CEAF_M$ | 29.11 | 29.11 | 29.11 | 34.96 | 34.96 | 34.96 | 28.78 | 28.78 | 28.78 |
| | $CEAF_E$ | 49.33 | 16.36 | 24.58 | 32.06 | 27.16 | 29.41 | 15.92 | 24.02 | 19.15 |
| | BLANC | 50.05 | 51.54 | 48.25 | 52.60 | 53.66 | 52.94 | 52.09 | 51.79 | 51.89 |
| | TOTAL | | | 30.54 | | | 40.10 | | | 37.53 |
| RuleH | MD | 7.35 | 48.95 | 12.78 | 57.09 | 57.57 | 57.33 | 71.80 | 65.37 | 68.44 |
| | MUC | 3.41 | 27.11 | 6.06 | 43.80 | 42.30 | 43.03 | 59.31 | 58.90 | 59.10 |
| | $B^3$ | 33.10 | 93.43 | 48.88 | 61.49 | 59.08 | 60.26 | 51.39 | 62.79 | 56.52 |
| | $CEAF_M$ | 29.79 | 29.79 | 29.79 | 42.55 | 42.55 | 42.55 | 40.59 | 40.59 | 40.59 |
| | $CEAF_E$ | 49.04 | 17.07 | 25.32 | 32.90 | 34.24 | 33.56 | 24.83 | 25.16 | 25.00 |
| | BLANC | 50.22 | 54.74 | 48.66 | 63.94 | 61.59 | 62.61 | 58.93 | 68.44 | 59.83 |
| | TOTAL | | | 31.74 | | | **48.40** | | | 48.21 |
| MLH | MD | 8.76 | 61.53 | 15.34 | 56.97 | 58.59 | 57.76 | 72.12 | 65.37 | 68.58 |
| | MUC | 4.05 | 35.18 | 7.26 | 42.51 | 42.10 | 42.30 | 59.54 | 58.90 | 59.22 |
| | $B^3$ | 31.90 | 94.26 | 47.66 | 59.05 | 59.56 | 59.30 | 51.57 | 62.67 | 56.58 |
| | $CEAF_M$ | 30.41 | 30.41 | 30.41 | 41.38 | 41.38 | 41.38 | 40.65 | 40.65 | 40.65 |
| | $CEAF_E$ | 52.28 | 17.28 | 25.98 | 33.40 | 33.77 | 33.58 | 24.78 | 25.29 | 25.03 |
| | BLANC | 50.37 | 60.43 | 48.83 | 59.09 | 59.44 | 59.26 | 58.92 | 68.38 | 59.81 |
| | TOTAL | | | **32.03** | | | 47.16 | | | **48.26** |
| gold heads | MD | 13.30 | 51.51 | 21.14 | 57.09 | 58.49 | 57.78 | 71.66 | 65.94 | 68.68 |
| | MUC | 4.69 | 20.18 | 7.61 | 42.83 | 42.28 | 42.56 | 58.61 | 58.90 | 58.76 |
| | $B^3$ | 36.24 | 87.14 | 51.19 | 59.40 | 59.54 | 59.47 | 49.33 | 62.52 | 55.15 |
| | $CEAF_M$ | 30.87 | 30.87 | 30.87 | 41.65 | 41.65 | 41.65 | 39.37 | 39.37 | 39.37 |
| | $CEAF_E$ | 46.06 | 18.87 | 26.77 | 33.48 | 33.97 | 33.72 | 24.77 | 24.54 | 24.60 |
| | BLANC | 50.26 | 52.50 | 49.09 | 59.28 | 59.50 | 59.38 | 58.08 | 67.40 | 58.51 |
| | TOTAL | | | 33.11 | | | 47.36 | | | 47.28 |

Table 3: MHD performance of HeuristicH, RuleH and MLH compared to the use of *gold* heads.

32.03 for Arabic and 48.26 for Chinese. For English, RuleH is marginally better than the MLH approach. This mirrors the performance of both methods in the intrinsic evaluation. Moreover, the performance of the system when given gold mention heads for this language is 47.36, which is only 0.2 percent points higher than MLH's performance. This shows that the latter approach already achieves a close to optimal performance.

The results of this experiment show that improvements in MHD translate directly into improvements of the overall CR system. Since the ML approach outperforms the rule-based approach for two languages, we can conclude that MLH is highly competitive for MHD in a MCR context, as it is language independent in that it does not require any language specific knowledge or annotation layers, apart from POS information and a small data set annotated for heads. Note also that the RuleH total scores for English and Chinese as well as the MLH total score for Chinese are higher than the respective values given *gold standard* heads. This is due to an increased recall across the different metrics.

## 6 Conclusion and Future Work

We propose a machine learning approach to mention head detection in the context of multilingual coreference resolution. We conducted an in-depth intrinsic and extrinsic evaluation of the method and compared it to a heuristic and a language dependent rule-based approach, generally used in CR systems. Our results show that the ML approach is language independent, given a small annotated set, and that it performs competitively in a multilingual setting.

The proposed ML method for MHD includes a basic set of language independent features. Like any ML approach, features are very important to the overall performance of the learner. For this reason, one very promising direction of further investigation is the thorough evaluation and extension of the feature set used for classification. In order to keep the language independent nature of MLH, only language independent features should be added to the current set of 14 values.

As discussed in section 4, the MLH approach does not control the number of heads allowed per mention. Thus, a possible improvement of this method can be achieved by an additional restriction on the number of heads allowed per phrase that is bound by the type of NP and the existence of coordinating conjunctions used in the phrase.

## Acknowledgments

We are grateful to the organizers of the shared task for providing the data sets and to Marwa Ragheb and Yuyin Hsu for the the mention head annotations in Arabic and Chinese.

## References

Amit Bagga and Breck Baldwin. 1998. Algorithms for scoring coreference chains. In *Proceedings of the LREC Workshop on Linguistic Coreference*, Granada, Spain.

Anders Björkelund and Richárd Farkas. 2012. Data-driven multilingual coreference resolution using resolver stacking. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 49–55, Jeju, Korea.

Chen Chen and Vincent Ng. 2012. Combining the best of two worlds: A hybrid approach to multilingual coreference resolution. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 56–63, Jeju, Korea.

Jinho D. Choi and Martha Palmer. 2010. Robust constituent-to-dependency conversion for English. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT)*, pages 55–66, Tartu, Estonia.

Noam Chomsky. 1970. Remarks on nominalization. In R. Jacobs and P. Rosenbaum, editors, *Reading in English Transformational Grammar*, pages 184–221. Ginn and Co., Waltham.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania, Philadelphia, PA, USA.

Walter Daelemans and Antal van den Bosch. 2005. *Memory Based Language Processing*. Cambridge University Press, Cambridge, UK.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2010. TiMBL: Tilburg Memory Based Learner, version 6.3, reference guide. Technical Report ILK 10-01, Induction of Linguistic Knowledge, Computational Linguistics, Tilburg University.

Ray Jackendoff. 1977. *X-Bar Syntax: A Study of Phrase Structure*. MIT Press, Cambridge.

Robert D. Levine and W. Detmar Meurers. 2006. Head-driven phrase structure grammar: Linguistic approach, formal foundations, and computational realization. In Keith Brown, editor, *Encyclopedia of Language and Linguistics*. Elsevier, 2nd ed. edition.

Xinxin Li, Xuan Wang, and Xingwei Liao. 2012. Simple maximum entropy models for multilingual coreference resolution. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 83–87, Jeju, Korea.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *HLT '05: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, Canada.

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

Sebastian Martschat, Jie Cai, Samuel Broscheit, Éva Mújdricza-Maydt, and Michael Strube. 2012. A multigraph model for coreference resolution. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 100–106, Jeju, Korea.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 104–111, Philadelphia, PA.

Vincent Ng. 2007. Shallow semantics for coreference resolution. In *Proceedings of the 20th International Conference on Artificial Intelligence (IJCAI'07)*, Hyderabad, India.

Sameer Pradhan, Lance Ramshaw, Mitchell Marcus, Martha Palmer, Ralph Weischedel, and Nianwen Xue. 2011. CoNLL-2011 shared task: Modeling unrestricted coreference in OntoNotes. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task*, pages 1–27, Portland, OR.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 1–40, Jeju, Korea.

Altaf Rahman and Vincent Ng. 2009. Supervised models for coreference resolution. In *Proceedings of EMNLP 2009*, pages 968–977, Singapore.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand Index for Coreference Evaluation. *NLE*, 17(4):485–510.

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.

Ivan A. Sag, Thomas Wasow, and Emily M. Bender. 2003. *Syntactic Theory: A Formal Introduction*. CSLI Publications, Stanford, 2 edition.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Honglin Sun and Daniel Jurafsky. 2004. Shallow Semantic Parsing of Chinese. In *North American Chapter of the ACL: Human Language Technologies (NAACL-HLT)*, pages 249–256, Boston, MA.

Olga Uryupina, Alessandro Moschitti, and Massimo Poesio. 2012. BART goes multilingual: The UniTN / Essex submission to the CoNLL-2012 shared task. In *Joint Conference on EMNLP and CoNLL-Shared Task*, pages 122–128, Jeju, Korea.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A Model-Theoretic Coreference Scoring Scheme. In *Proceedings of the 6th Message Understanding Conference*, Columbia, MD.

Yue Zhang and Stephen Clark. 2011. Syntactic Processing Using the Generalized Perceptron and Beam Search. *Computational Linguistics*, 37(1):105–151.

Desislava Zhekova and Sandra Kübler. 2010. UBIU: A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96–99, Uppsala, Sweden.