Unsupervised Ranking Model for Entity Coreference Resolution

Xuezhe Ma and **Zhengzhong Liu** and **Eduard Hovy**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{xuezhem, liu}@cs.cmu.edu, ehovy@cmu.edu

Abstract

Coreference resolution is one of the first stages in deep language understanding and its importance has been well recognized in the natural language processing community. In this paper, we propose a generative, unsupervised ranking model for entity coreference resolution by introducing resolution mode variables. Our unsupervised system achieves 58.44% F1 score of the CoNLL metric on the English data from the CoNLL-2012 shared task (Pradhan et al., 2012), outperforming the Stanford deterministic system (Lee et al., 2013) by 3.01%.

1 Introduction

Entity coreference resolution has become a critical component for many Natural Language Processing (NLP) tasks. Systems requiring deep language understanding, such as information extraction (Wellner et al., 2004), semantic event learning (Chambers and Jurafsky, 2008; Chambers and Jurafsky, 2009), and named entity linking (Durrett and Klein, 2014; Ji et al., 2014) all benefit from entity coreference information.

Entity coreference resolution is the task of identifying mentions (i.e., noun phrases) in a text or dialogue that refer to the same real-world entities. In recent years, several supervised entity coreference resolution systems have been proposed, which, according to Ng (2010), can be categorized into three classes — mention-pair models (McCarthy and Lehnert, 1995), entity-mention models (Yang et al., 2008a; Haghighi and Klein, 2010; Lee et al., 2011) and ranking models (Yang et al., 2008b; Durrett and Klein, 2013; Fernandes et al., 2014) — among which ranking models recently obtained state-of-the-art performance. However, the manually annotated corpora that these systems rely on are highly expensive to create, in particular when we want to build data for resource-poor languages (Ma and Xia, 2014). That makes unsupervised approaches, which only require unannotated text for training, a desirable solution to this problem.

Several unsupervised learning algorithms have been applied to coreference resolution. Haghighi and Klein (2007) presented a mention-pair nonparametric fully-generative Bayesian model for unsupervised coreference resolution. Based on this model, Ng (2008) probabilistically induced coreference partitions via EM clustering. Poon and Domingos (2008) proposed an entity-mention model that is able to perform joint inference across mentions by using Markov Logic. Unfortunately, these unsupervised systems' performance on accuracy significantly falls behind those of supervised systems, and are even worse than the deterministic rule-based systems. Furthermore, there is no previous work exploring the possibility of developing an unsupervised ranking model which achieved state-of-the-art performance under supervised settings for entity coreference resolution.

In this paper, we propose an unsupervised generative ranking model for entity coreference resolution. Our experimental results on the English data from the CoNLL-2012 shared task (Pradhan et al., 2012) show that our unsupervised system outperforms the Stanford deterministic system (Lee et al., 2013) by 3.01% absolute on the CoNLL official metric. The

Proceedings of NAACL-HLT 2016, pages 1012–1018,
San Diego, California, June 12-17, 2016.

contributions of this work are (i) proposing the first unsupervised ranking model for entity coreference resolution. (ii) giving empirical evaluations of this model on benchmark data sets. (iii) considerably narrowing the gap to supervised coreference resolution accuracy.

2 Unsupervised Ranking Model

2.1 Notations and Definitions

In the following, $D = \{m_0, m_1, \ldots, m_n\}$ represents a generic input document which is a sequence of coreference mentions, including the artificial root mention (denoted by m_0). The method to detect and extract these mentions is discussed later in Section 2.6. Let $C = \{c_1, c_2, \ldots, c_n\}$ denote the coreference assignment of a given document, where each mention m_i has an associated random variable c_i taking values in the set $\{0, i, \ldots, i-1\}$; this variable specifies m_i's selected antecedent ($c_i \in \{1, 2, \ldots, i-1\}$), or indicates that it begins a new coreference chain ($c_i = 0$).

2.2 Generative Ranking Model

The following is a straightforward way to build a generative model for coreference:

$$\begin{aligned} P(D, C) &= P(D|C)P(C) \\ &= \prod_{j=1}^{n} P(m_j|m_{c_j}) \prod_{j=1}^{n} P(c_j|j) \end{aligned} \quad (1)$$

where we factorize the probabilities $P(D|C)$ and $P(C)$ into each position j by adopting appropriate independence assumptions that given the coreference assignment c_j and corresponding coreferent mention m_{c_j}, the mention m_j is independent with other mentions in front of it. This independent assumption is similar to that in the IBM 1 model on machine translation (Brown et al., 1993), where it assumes that given the corresponding English word, the aligned foreign word is independent with other English and foreign words. We do not make any independent assumptions among different features (see Section 2.4 for details).

Inference in this model is efficient, because we can compute c_j separately for each mention:

$$c_j^* = \operatorname*{argmax}_{c_j} P(m_j|m_{c_j})P(c_j|j)$$

The model is a so-called ranking model because it is able to identify the most probable candidate antecedent given a mention to be resolved.

2.3 Resolution Mode Variables

According to previous work (Haghighi and Klein, 2009; Ratinov and Roth, 2012; Lee et al., 2013), antecedents are resolved by different categories of information for different mentions. For example, the Stanford system (Lee et al., 2013) uses string-matching sieves to link two mentions with similar text and precise-construct sieve to link two mentions which satisfy special syntactic or semantic relations such as apposition or acronym. Motivated by this, we introduce resolution mode variables $\Pi = \{\pi_1, \ldots, \pi_n\}$, where for each mention j the variable $\pi_j \in \{str, prec, attr\}$ indicates in which mode the mention should be resolved. In our model, we define three resolution modes — string-matching (*str*), precise-construct (*prec*), and attribute-matching (*attr*) — and Π is deterministic when D is given (i.e. $P(\Pi|D)$ is a point distribution). We determine π_j for each mention m_j in the following way:

- $\pi_j = str$, if there exists a mention $m_i, i < j$ such that the two mentions satisfy the *String Match* sieve, the *Relaxed String Match* sieve, or the *Strict Head Match A* sieve in the Stanford multi-sieve system (Lee et al., 2013).

- $\pi_j = prec$, if there exists a mention $m_i, i < j$ such that the two mentions satisfy the *Speaker Identification* sieve, or the *Precise Constructs* sieve.

- $\pi_j = attr$, if there is no mention $m_i, i < j$ satisfies the above two conditions.

Now, we can extend the generative model in Eq. 1 to:

$$\begin{aligned} & P(D, C) = P(D, C, \Pi) \\ = & \prod_{j=1}^{n} P(m_j|m_{c_j}, \pi_j)P(c_j|\pi_j, j)P(\pi_j|j) \end{aligned}$$

where we define $P(\pi_j|j)$ to be uniform distribution. We model $P(m_j|m_{c_j}, \pi_j)$ and $P(c_j|\pi_j, j)$ in the fol-

Mode π	Feature	Description
prec	Mention Type	the type of a mention. We use three mention types: $Proper, Nominal, Pronoun$
str	Mention Type	the same as the mention type feature under *prec* mode.
	Exact Match	boolean feature corresponding to String Match sieve in Stanford system.
	Relaxed Match	boolean feature corresponding to Relaxed String Match sieve in Stanford system.
	Head Match	boolean feature corresponding to Strict Head Match A sieve in Stanford system.
attr	Mention Type	the same as the mention type feature under *prec* mode.
	Number	the number of a mention similarly derived from Lee et al. (2013).
	Gender	the gender of a mention from Bergsma and Lin (2006) and Ji and Lin (2009).
	Person	the person attribute from Lee et al. (2013). We assign person attributes to all mentions, not only pronouns.
	Animacy	the animacy attribute same as Lee et al. (2013).
	Semantic Class	semantic classes derived from WordNet (Soon et al., 2001).
	Distance	sentence distance between the two mentions. This feature is for parameter $q(k\|j,\pi)$

Table 1: Feature set for representing a mention under different resolution modes. The *Distance* feature is for parameter q, while all other features are for parameter t.

Algorithm 1: Learning Model with EM

Initialization: Initialize $\theta_0 = \{t_0, q_0\}$
for $t = 0$ ***to*** T **do**
  set all counts $c(\ldots) = 0$
  for *each document* D **do**
    for $j = 1$ ***to*** n **do**
      for $k = 0$ ***to*** $j-1$ **do**
        $L_{jk} = \frac{t(m_j|m_k,\pi_j)q(k|\pi_j,j)}{\sum_{i=0}^{j-1} t(m_j|m_i,\pi_j)q(i|\pi_j,j)}$
        $c(m_j, m_k, \pi_j)$ += L_{jk}
        $c(m_k, \pi_j)$ += L_{jk}
        $c(k, j, \pi_j)$ += L_{jk}
        $c(j, \pi_j)$ += L_{jk}
  // Recalculate the parameters
  $t(m|m', \pi) = \frac{c(m,m',\pi)}{c(m',\pi)}$
  $q(k, j, \pi) = \frac{c(k,j,\pi)}{c(j,\pi)}$

lowing way:

$$P(m_j|m_{c_j}, \pi_j) = t(m_j|m_{c_j}, \pi_j)$$
$$P(c_j|\pi_j, j) = \begin{cases} q(c_j|\pi_j, j) & \pi_j = attr \\ \frac{1}{j} & \text{otherwise} \end{cases}$$

where $\theta = \{t, q\}$ are parameters of our model. Note that in the attribute-matching mode ($\pi_j = attr$) we model $P(c_j|\pi_j, j)$ with parameter q, while in the other two modes, we use the uniform distribution. It makes sense because the position information is important for coreference resolved by matching attributes of two mentions such as resolving pronoun coreference, but not that important for those resolved by matching text or special relations like two mentions referring the same person and matching by the name.

Corpora	# Doc	# Sent	# Word	# Entity	# Mention
Gigaword	3.6M	75.4M	1.6B	-	-
ON-Dev	343	9,142	160K	4,546	19,156
ON-Test	348	9,615	170K	4,532	19,764

Table 2: Corpora statistics. "ON-Dev" and "ON-Test" are the development and testing sets of the OntoNotes corpus.

2.4 Features

In this section, we describe the features we use to represent mentions. Specifically, as shown in Table 1, we use different features under different resolution modes. It should be noted that only the *Distance* feature is designed for parameter q, all other features are designed for parameter t.

2.5 Model Learning

For model learning, we run EM algorithm (Dempster et al., 1977) on our Model, treating D as observed data and C as latent variables. We run EM with 10 iterations and select the parameters achieving the best performance on the development data. Each iteration takes around 12 hours with 10 CPUs parallelly. The best parameters appear at around the 5th iteration, according to our experiments.The detailed derivation of the learning algorithm is shown in Appendix A. The pseudo-code is shown is Algorithm 1. We use uniform initialization for all the parameters in our model.

Several previous work has attempted to use EM for entity coreference resolution. Cherry and Bergsma (2005) and Charniak and Elsner (2009) applied EM for pronoun anaphora resolution. Ng (2008) probabilistically induced coreference partitions via EM clustering. Recently, Moosavi and Strube (2014) proposed an unsupervised model uti-

	CoNLL'12 English development data						CoNLL'12 English test data					
	MUC	B^3	$CEAF_m$	$CEAF_e$	Blanc	CoNLL	MUC	B^3	$CEAF_m$	$CEAF_e$	Blanc	CoNLL
MIR	65.39	54.89	–	51.36	–	57.21	64.64	52.52	–	49.11	–	55.42
Stanford	64.96	54.49	59.39	51.24	56.03	56.90	64.71	52.26	56.01	49.32	53.92	55.43
Multigraph	66.22	56.41	60.87	52.61	58.15	58.41	65.41	54.38	58.60	50.21	56.03	56.67
Our Model	**67.89**	**57.83**	**62.11**	**53.76**	**60.58**	**59.83**	**67.69**	**55.86**	**59.66**	**51.75**	**57.78**	**58.44**
IMS	67.15	55.19	58.86	50.94	56.22	57.76	67.58	54.47	58.17	50.21	55.41	57.42
Latent-Tree	69.46	57.83	–	54.43	–	60.57	70.51	57.58	–	53.86	–	60.65
Berkeley	70.44	59.10	–	55.57	–	61.71	70.62	58.20	–	54.80	–	61.21
LaSO	70.74	60.03	65.01	56.80	–	62.52	70.72	58.58	63.45	59.40	–	61.63
Latent-Strc	72.11	60.74	–	57.72	–	63.52	72.17	59.58	–	55.67	–	62.47
Model-Stack	72.59	61.98	–	57.58	–	64.05	72.59	60.44	–	56.02	–	63.02
Non-Linear	72.74	61.77	–	58.63	–	64.38	72.60	60.52	–	57.05	–	63.39

Table 3: F1 scores of different evaluation metrics for our model, together with two deterministic systems and one unsupervised system as baseline (above the dashed line) and seven supervised systems (below the dashed line) for comparison on CoNLL 2012 development and test datasets.

lizing the most informative relations and achieved competitive performance with the Stanford system.

2.6 Mention Detection

The basic rules we used to detect mentions are similar to those of Lee et al. (2013), except that their system uses a set of filtering rules designed to discard instances of pleonastic *it*, partitives, certain quantified noun phrases and other spurious mentions. Our system keeps partitives, quantified noun phrases and *bare NP* mentions, but discards pleonastic *it* and other spurious mentions.

3 Experiments

3.1 Experimental Setup

Datasets. Due to the availability of readily parsed data, we select the APW and NYT sections of Gigaword Corpus (years 1994-2010) (Parker et al., 2011) to train the model. Following previous work (Chambers and Jurafsky, 2008), we remove duplicated documents and the documents which include fewer than 3 sentences. The development and test data are the English data from the CoNLL-2012 shared task (Pradhan et al., 2012), which is derived from the OntoNotes corpus (Hovy et al., 2006). The corpora statistics are shown in Table 2. Our system is evaluated with automatically extracted mentions on the version of the data with automatic preprocessing information (e.g., predicted parse trees).

Evaluation Metrics. We evaluate our model on three measures widely used in the literature: MUC (Vilain et al., 1995), B^3 (Bagga and Baldwin, 1998), and Entity-based CEAF ($CEAF_e$) (Luo, 2005). In addition, we also report results on another two popular metrics: Mention-based CEAF ($CEAF_m$) and BLANC (Recasens and Hovy, 2011). All the results are given by the latest version of CoNLL-2012 scorer [1]

3.2 Results and Comparison

Table 3 illustrates the results of our model together as baseline with two deterministic systems, namely **Stanford**: the Stanford system (Lee et al., 2011) and **Multigraph**: the unsupervised multigraph system (Martschat, 2013), and one unsupervised system, namely **MIR**: the unsupervised system using most informative relations (Moosavi and Strube, 2014). Our model outperforms the three baseline systems on all the evaluation metrics. Specifically, our model achieves improvements of 2.93% and 3.01% on CoNLL F1 score over the Stanford system, the winner of the CoNLL 2011 shared task, on the CoNLL 2012 development and test sets, respectively. The improvements on CoNLL F1 score over the Multigraph model are 1.41% and 1.77% on the development and test sets, respectively. Comparing

[1] http://conll.cemantix.org/2012/software.html

with the MIR model, we obtain significant improvements of 2.62% and 3.02% on CoNLL F1 score.

To make a thorough empirical comparison with previous studies, Table 3 (below the dashed line) also shows the results of some state-of-the-art supervised coreference resolution systems — **IMS**: the second best system in the CoNLL 2012 shared task (Björkelund and Farkas, 2012); **Latent-Tree**: the latent tree model (Fernandes et al., 2012) obtaining the best results in the shared task; **Berkeley**: the Berkeley system with the final feature set (Durrett and Klein, 2013); **LaSO**: the structured perceptron system with non-local features (Björkelund and Kuhn, 2014); **Latent-Strc**: the latent structure system (Martschat and Strube, 2015); **Model-Stack**: the entity-centric system with model stacking (Clark and Manning, 2015); and **Non-Linear**: the non-linear mention-ranking model with feature representations (Wiseman et al., 2015). Our unsupervised ranking model outperforms the supervised IMS system by 1.02% on the CoNLL F1 score, and achieves competitive performance with the latent tree model. Moreover, our approach considerably narrows the gap to other supervised systems listed in Table 3.

4 Conclusion

We proposed a new generative, unsupervised ranking model for entity coreference resolution into which we introduced resolution mode variables to distinguish mentions resolved by different categories of information. Experimental results on the data from CoNLL-2012 shared task show that our system significantly improves the accuracy on different evaluation metrics over the baseline systems.

One possible direction for future work is to differentiate more resolution modes. Another one is to add more precise or even event-based features to improve the model's performance.

Acknowledgements

This research was supported in part by DARPA grant FA8750-12-2-0342 funded under the DEFT program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

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*, volume 1, pages 563–566. Citeseer.

Shane Bergsma and Dekang Lin. 2006. Bootstrapping path-based pronoun resolution. In *Proceedings of ACL-2006*, pages 33–40, Sydney, Australia, July. Association for Computational Linguistics.

Anders Björkelund and Richárd Farkas. 2012. Data-driven multilingual coreference resolution using resolver stacking. In *Proceedings of EMNLP-CoNLL-2012 - Shared Task*, pages 49–55, Jeju Island, Korea, July. Association for Computational Linguistics.

Anders Björkelund and Jonas Kuhn. 2014. Learning structured perceptrons for coreference resolution with latent antecedents and non-local features. In *Proceedings of ACL-2014*, pages 47–57, Baltimore, Maryland, June. Association for Computational Linguistics.

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-2008: HLT*, pages 789–797, Columbus, Ohio, June. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of ACL-2009*, pages 602–610, Suntec, Singapore, August. Association for Computational Linguistics.

Eugene Charniak and Micha Elsner. 2009. EM works for pronoun anaphora resolution. In *Proceedings of EACL 2009*, pages 148–156, Athens, Greece, March.

Colin Cherry and Shane Bergsma. 2005. An Expectation Maximization approach to pronoun resolution. In *Proceedings of CoNLL-2005*, pages 88–95, Ann Arbor, Michigan, June.

Kevin Clark and Christopher D. Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proceedings of ACL-IJCNLP-2015*, pages 1405–1415, Beijing, China, July. Association for Computational Linguistics.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of EMNLP-2013*, pages 1971–1982, Seattle,

Washington, USA, October. Association for Computational Linguistics.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. In *Proceedings of the Transactions of the Association for Computational Linguistics*.

Eraldo Fernandes, Cícero dos Santos, and Ruy Milidiú. 2012. Latent structure perceptron with feature induction for unrestricted coreference resolution. In *Proceedings of EMNLP-CoNLL-2012 - Shared Task*, pages 41–48, Jeju Island, Korea, July. Association for Computational Linguistics.

Eraldo Rezende Fernandes, Cícero Nogueira dos Santos, and Ruy Luiz Milidiú. 2014. Latent trees for coreference resolution. *Computational Linguistics*.

Aria Haghighi and Dan Klein. 2007. Unsupervised coreference resolution in a nonparametric bayesian model. In *Proceedings of ACL-2007*, pages 848–855, Prague, Czech Republic, June. Association for Computational Linguistics.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of EMNLP-2009*, pages 1152–1161, Singapore, August. Association for Computational Linguistics.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Proceedings of NAACL-2010*, pages 385–393, Los Angeles, California, June. Association for Computational Linguistics.

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

Heng Ji and Dekang Lin. 2009. Gender and animacy knowledge discovery from web-scale n-grams for unsupervised person mention detection. In *Proceedings of PACLIC-2009*, pages 220–229.

Heng Ji, HT Dang, J Nothman, and B Hachey. 2014. Overview of tac-kbp2014 entity discovery and linking tasks. In *Proc. Text Analysis Conference (TAC2014)*.

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 CoNLL-2011: Shared Task*, pages 28–34, Portland, Oregon, USA, June. Association for Computational Linguistics.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Comput. Linguist.*, 39(4):885–916, December.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of EMNLP-2005*, pages 25–32, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Xuezhe Ma and Fei Xia. 2014. Unsupervised dependency parsing with transferring distribution via parallel guidance and entropy regularization. In *Proceedings of ACL-2014*, pages 1337–1348, Baltimore, Maryland, June.

Sebastian Martschat and Michael Strube. 2015. Latent structures for coreference resolution. *Transactions of the Association for Computational Linguistics*, 3:405–418.

Sebastian Martschat. 2013. Multigraph clustering for unsupervised coreference resolution. In *ACL-2013: Student Research Workshop*, pages 81–88, Sofia, Bulgaria, August. Association for Computational Linguistics.

Joseph F McCarthy and Wendy G Lehnert. 1995. Using decision trees for conference resolution. In *Proceedings of IJCAI-1995*, pages 1050–1055. Morgan Kaufmann Publishers Inc.

Nafise Sadat Moosavi and Michael Strube. 2014. Unsupervised coreference resolution by utilizing the most informative relations. In *Proceedings of COLING-2014*, pages 644–655.

Vincent Ng. 2008. Unsupervised models for coreference resolution. In *Proceedings of EMNLP-2008*, pages 640–649, Honolulu, Hawaii, October. Association for Computational Linguistics.

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

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition. *Linguistic Data Consortium, LDC2011T07*.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with Markov Logic. In *Proceedings of EMNLP-2008*, pages 650–659, Honolulu, Hawaii, October. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Proceedings of EMNLP-CoNLL-2012 - Shared Task*, pages 1–40, Jeju Island, Korea, July. Association for Computational Linguistics.

Lev Ratinov and Dan Roth. 2012. Learning-based multi-sieve co-reference resolution with knowledge. In *Proceedings of EMNLP-CoNLL-2012*, pages 1234–1244, Jeju Island, Korea, July. Association for Computational Linguistics.

Marta Recasens and Eduard Hovy. 2011. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(04):485–510.

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.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52. Association for Computational Linguistics.

Ben Wellner, Andrew McCallum, Fuchun Peng, and Michael Hay. 2004. An integrated, conditional model of information extraction and coreference with application to citation matching. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 593–601. AUAI Press.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proceedings of ACL-IJCNLP-2015*, pages 1416–1426, Beijing, China, July. Association for Computational Linguistics.

Xiaofeng Yang, Jian Su, Jun Lang, Chew Lim Tan, Ting Liu, and Sheng Li. 2008a. An entity-mention model for coreference resolution with inductive logic programming. In *Proceedings of ACL-2008*, pages 843–851.

Xiaofeng Yang, Jian Su, and Chew Lim Tan. 2008b. A twin-candidate model for learning-based anaphora resolution. *Computational Linguistics*, 34(3):327–356.

Appendix A. Derivation of Model Learning

Formally, we iteratively estimate the model parameters θ, employing the following EM algorithm:

E-step: Compute the posterior probabilities of C, $P(C|D;\theta)$, based on the current θ.

M-step: Calculate the new θ' that maximizes the expected complete log likelihood, $E_{P(C|D;\theta)}[\log P(D,C;\theta')]$

For simplicity, we denote:

$$\begin{aligned} P(C|D;\theta) &= \tilde{P}(C|D) \\ P(C|D;\theta') &= P(C|D) \end{aligned}$$

In addition, we use $\tau(\pi_j|j)$ to denote the probability $P(\pi_j|j)$ which is uniform distribution in our model. Moreover, we use the following notation for convenience:

$$\theta(m_j, m_k, j, k, \pi_j) = t(m_j|m_k, \pi_j)q(k|\pi_j, j)\tau(\pi_j|j)$$

Then, we have

$$\begin{aligned} & E_{\tilde{P}(c|D)}[\log P(D,C)] \\ =& \sum_C \tilde{P}(C|D) \log P(D,C) \\ =& \sum_C \tilde{P}(C|D)\big(\sum_{j=1}^{n} \log t(m_j|m_{c_j}, \pi_j) + \log q(c_j|\pi_j, j) + \log \tau(\pi_j|j)\big) \\ =& \sum_{j=1}^{n}\sum_{k=0}^{j-1} L_{jk}\big(\log t(m_j|m_k, \pi_j) + \log q(k|\pi_j, j) + \log \tau(\pi_j|j)\big) \end{aligned}$$

Then the parameters t and q that maximize $E_{\tilde{P}(c|D)}[\log P(D,C)]$ satisfy that

$$\begin{aligned} t(m_j|m_k, \pi_j) &= \frac{L_{jk}}{\sum_{i=1}^{n} L_{ik}} \\ q(k|\pi_j, j) &= \frac{L_{jk}}{\sum_{i=0}^{j-1} L_{ji}} \end{aligned}$$

where L_{jk} can be calculated by

$$\begin{aligned} L_{jk} &= \sum_{C, c_j=k} \tilde{P}(C|D) = \frac{\sum_{C,c_j=k} \tilde{P}(C,D)}{\sum_C \tilde{P}(C,D)} \\ &= \frac{\sum_{C,c_j=k}\prod_{i=1}^{n} \tilde{\theta}(m_i, m_{c_i}, c_i, i, \pi_i)}{\sum_C \prod_{i=1}^{n} \tilde{\theta}(m_i, m_{c_i}, c_i, i, \pi_i)} \\ &= \frac{\tilde{\theta}(m_j, m_k, k, j, \pi_j)\sum_{C(-j)} \tilde{P}(C(-j)|D)}{\sum_{i=0}^{j-1} \tilde{\theta}(m_j, m_i, i, j, \pi_j)\sum_{C(-j)} \tilde{P}(C(-j)|D)} \\ &= \frac{\tilde{\theta}(m_j, m_k, k, j, \pi_j)}{\sum_{i=0}^{j-1} \tilde{\theta}(m_j, m_i, i, j, \pi_j)} \\ &= \frac{\tilde{t}(m_j|m_k, \pi_j)\tilde{q}(k|\pi_j, j)\tilde{\tau}(\pi_j|j)}{\sum_{i=0}^{j-1} \tilde{t}(m_j|m_i, \pi_j)\tilde{q}(i|\pi_j, j)\tilde{\tau}(\pi_j|j)} \\ &= \frac{\tilde{t}(m_j|m_k, \pi_j)\tilde{q}(k|\pi_j, j)}{\sum_{i=0}^{j-1} \tilde{t}(m_j|m_i, \pi_j)\tilde{q}(i|\pi_j, j)} \end{aligned}$$

where $C(-j) = \{c_1, \ldots, c_{j-1}, c_{j+1}, \ldots, c_n\}$. The above derivations correspond to the learning algorithm in Algorithm 1.