Appendix I. Extended Related Works

Coreference resolution research steered from rule-based towards machine learning approaches in the 1990s. This transition emerged with the availability of annotated coreference corpora[10-13] for semantic categories like organizations, persons, locations, dates, times, money, or percentages.

McCarthy and Lehnert's supervised machine learning system, RESOLVE,[3] resolved coreference of NPs in four steps: pair creation, feature set determination, learning, and clustering. RESOLVE focused on four semantic types: organizations, facilities, persons, and products-or-services. For each semantic type, RESOLVE created all pairs of ordered markables. It characterized each pair by eight features, e.g., whether the pair referred to a business joint venture, whether the pair contained a personal name, and whether the markables in the pair presented overlapping tokens (i.e., lexical words), shared a common NP, or originated from the same sentence. RESOLVE applied the C4.5[14] decision tree to classify pairs as coreferent or not. It applied the "aggressive merge"[3] clustering algorithm to create chains from coreferent pairs. RESOLVE represented state-of-the-art results at its time. It achieved on the Message Understanding Conference-5 (MUC-5) corpus an average F-measure of .858 over coreference chains.

Soon et al.[15] extended RESOLVE to pronouns. They expanded RESOLVE's feature set, modified RESOLVE's pair creation step, and applied the Closest-Link algorithm for generating chains. Their pair creation step only produced positive training pairs from neighboring markables in a chain. In chain *A-B-C-D*, they selected as positive training pairs *A-B, B-C,* and *C-D*. If the text contained the ordered set of non-coreferent markables *a, b*, and *c* between the coreferent markables *A* and *B*, then negative pairs *a-B, b-B*, and *c-B* were also generated. Their clustering algorithm linked each markable to its closest satisfactory antecedent. A markable *A* was assigned to a chain *K* if the closest preceding markable that was classified as coreferent to *A* was a member of chain *K*. Overall, Soon et al.'s algorithm gave an F-measure of .626 on MUC-6, compared to RESOLVE's F-measure of .472 on the same corpus. It found the token overlap to be the single most informative feature with an F-measure of .539. Versley et al.[16] implemented Soon et al.'s algorithm in the Beautiful Anaphora Resolution Toolkit (BART). On the MUC-7 corpus, BART gave a best precision of .741 using the support vector machine (SVM) linear classifier, and a best recall of .563 using the maximum entropy classifier.

Ng and Cardie[17] improved Soon et al.'s system by modifying their feature set and the pair creation algorithm. They observed that the token overlap between markables was more informative for coreference resolution on proper names than on pronouns; they created a token overlap feature for each category of pronouns, proper names, and non-pronominal NPs. Ng and Cardie complemented their system with a

Best-Link clustering algorithm which linked each markable with its most likely antecedent, and chains the selected markables together based on transitivity. Ng and Cardie's system evaluated on MUC-6 to F-measure .704 and on MUC-7 to F-measure .634.

Yang et al.[18] extended RESOLVE's feature set in order to examine how the different methods of measuring the token overlap affected coreference resolution. Their system achieved a best recall of .714 and best precision of .697. Castano et al.[19] deviated from RESOLVE's framework for sortal and pronominal coreference resolution on Medline abstracts. They used the UMLS Metathesaurus[20] and MetaMap to identify biomedical markables and their semantic types. They achieved a precision of .733 and a recall of .700.

Son et al.[21] studied co-reference of findings of lung masses in radiology documents. They used a probabilistically guided model that incorporated domain knowledge (e.g., mass location, quantity, size, calcification pattern). Their system achieved .672 MUC F-measure. Yangy et al.[22] argued that coreference chains could be more informative than individual NPs for coreference resolution. Their system outperformed those of Soon et al. and Castano et al. on MEDLINE abstracts, with an F-measure of .817.

Stoyanov et al.[23] developed $\text{RECONCILE}_{\text{ACL09}}$, which they modeled after the state-of-the art system of Ng and Cardie. They used a set of 76 features proven successful in the literature, and applied the perceptron learning algorithm for classification and a single-link algorithm for clustering. When evaluated on the MUC-6 and MUC-7 corpora, $\text{RECONCILE}_{\text{ACL09}}$[24] outperformed the Soon et al. and Ng and Cardie systems with a .712 MUC F-measure on MUC-6 and .629 MUC F-measure on MUC-7.

Appendix II. Evaluation of pair classification

We evaluated performance on classification of pairs as coreferent or non-coreferent using precision, recall, and F-measure. Precision computed the number of correct pair classifications in a class (true positives) divided by the number of total pair classifications (true positives and false positives). Recall represented the number of true positive pairs in a class divided by the total number of pairs in the class (true positives and false negatives). F-measure was the harmonic mean of precision and recall. MCORES had a pair classification F-measure across all markables of .824 on the per-entity runs and .603 on the per-corpus runs (see Table 6). The best pair classification F-measure occurred for person pairs (F-measure .898 per-entity runs and .649 per-corpus runs) and the lowest occurred for problem pairs (F-measure .776) on the per-entity runs, and for tests on the per-corpus runs (F-measure .420).

Appendix III. Evaluation metrics for coreference resolution

MUC metrics[29] assessed the minimal number of pairs that needed to be added or taken away from a chain in order for it to match the gold standard. Links that needed to be added were treated as false negatives; links that needed to be removed were false positives.

Let K represent all coreference chains in the gold standard, and R the chains generated by the system on the markables in K. Given chains k and r from K and R, respectively, MUC recall and precision of R were:

$$recall = \frac{\sum_{k}(|k| - m(k,R))}{\sum_{k}(|k| - 1)}$$

$$precision = \frac{\sum_{k}(|r| - m(k,K))}{\sum_{k}(|r| - 1)}$$

Where m(r,K) represented the number of coreference chains in K that intersected the chain r.

The MUC F-measure of chains was given by:

$$F - measure = \frac{2 * recall * precision}{recall + precision}$$

B^3 metrics[30] evaluated chains by measuring the overlap between the predicted chains and the gold standard.

Let C be a corpus, d a document, and m a markable within the document, where the total number of markables in C was N. Let G_m be the gold standard coreference chain that contains m, S_m the system generated chain that contains m, and O_m the intersection of G_m and S_m. B^3 recall and precision were:

$$\text{recall} = \frac{1}{N}\sum_{d \in C}\sum_{m \in d}\frac{|O_m|}{|G_m|}$$

$$precision = \frac{1}{N}\sum_{d \in C}\sum_{m \in d}\frac{|O_m|}{|S_m|}$$

The B^3 F-measure was identical to MUC F-measure.

CEAF[31] metrics aligned chains in the system response and gold standard and used the best mapping to compute the CEAF precision recall and F-measure. The chain alignment was computed based on a similarity score, which could be either markable- or chain-based. There were two variants for the chain-based similarity score, ϕ_3 and ϕ_4. We employed ϕ_4, unless otherwise specified.

$$\phi_3(K_i, R_j) = |K_i \cap R_j|$$
$$\phi_4(K_i, R_j) = \frac{2|K_i \cap R_j|}{|K_i| + |R_j|}$$

The CEAF precision and recall depended on the result of the alignment which had the best total similarity score (denoted as $\Phi(g^*)$):

$$\mathrm{Precision} = \frac{\Phi(g^*)}{\sum_i \phi(R_i, R_i)}$$
$$\mathrm{Recall} = \frac{\Phi(g^*)}{\sum_i \phi(K_i, K_i)}$$

BLANC[32] metrics defined the notions of coreference decisions and correctness decisions. The coreference decisions (i.e., coreference link (c) and non-coreference link (n)) were made by the coreference system. The correctness decisions were made by the evaluator and comprised of right link (r), which had the same value of either coreference or non-coreference in the gold standard and system response, and wrong link (w), which had different values in the gold standard and system response. The BLANC metric computed precision and recall for the coreference (P_c, R_c) and non-coreferent (P_n, R_n) links. The final precision and recall were an unweighted average of the earlier results.

$$P_c = \frac{rc}{rc + wc} \quad R_c = \frac{rc}{rc + wn}$$
$$P_n = \frac{rn}{rn + wn} \quad R_n = \frac{rn}{rn + wc}$$

Appendix IV. Significance testing

We used the approximate randomization test[34] to assess whether two system outputs were significantly different from each other. For a pair of outputs A and B from two different systems, we computed the unweighted average F-measures f_A for output A and f_B for output B, as well as the absolute difference in performance $f=|f_A-f_B|$. Given A's set of entries of length j, and B's set of entries of length k, we created superset C, of size j+k, by joining A and B's system entries. For each iteration *i* up to N iterations, we selected from C j entries randomly and without resampling and created the pseudoset of entries A_i. The remainder of k elements in C created the pseudoset of entries B_i. We computed the unweighted F-measures f_A' for A_i and f_B' for B_i, and noted the absolute difference between them ($f_i = |f_A' - f_B'|$). We computed Nt, the number of times that $f_i - f >= 0$ and calculated the p value between A and B as p= (Nt +1)/(N+1). We ran significance tests with the number of iterations N=3000 and alpha=0.05.

We applied the Bonferroni correction to adjust for multiple comparisons. Our corrected alpha was set to .00045 (obtained by dividing 0.05 with 111), for 111 comparisons (see Table 5).

Table 2: Number of markables per coreference chain: min, max, and average

	Persons	**Problems**	**Treatments**	**Tests**	**Across all chains**
Min	2	2	2	2	2
Max	149	17	20	10	149
Average	13.236	2.932	2.638	2.317	5.280

Table 6: Precision, recall, and F-measure of MCORES for pair classification

	Per-entity Runs					Per-corpus Runs				
	Persons	**Problems**	**Treatments**	**Tests**	**Across all markables**	**Persons**	**Problems**	**Treatments**	**Tests**	**Across all markables**
Precision	.902	.703	.754	.754	.778	.536	.429	.485	.313	.484
Recall	.894	.866	.911	.849	.880	.821	.833	.794	.637	.799
F-measure	.898	.776	.825	.799	.824	.649	.566	.602	.420	.603

Table 7: MUC, B^3, CEAF, and BLANC F-measures, and the unweighted averages of MUC, B^3, CEAF, and BLANC F-measures for MCORES and baseline

		Per-entity Runs					Per-corpus Runs				
		Persons	**Problems**	**Treatments**	**Tests**	**Across all markables**	**Persons**	**Problems**	**Treatments**	**Tests**	**Across all markables**
Precision											
MUC	MCORES	.898	.630	.689	.505	.797	.500	.409	.465	.364	.497
	Baseline	.587	.614	.687	.532	.604	.396	.426	.496	.366	.429
B^3	MCORES	.944	.990	.995	.998	.990	.957	.988	.987	.980	.984
	Baseline	.981	.991	.991	.996	.991	.979	.992	.992	.990	.991
CEAF	MCORES	.667	.889	.912	.970	.878	.252	.809	.837	.947	.835
	Baseline	.275	.882	.913	.974	.791	.214	.815	.848	.949	.820
BLANC	MCORES	.925	.840	.905	.853	.922	.846	.793	.701	.636	.751
	Baseline	.907	.839	.811	.789	.895	.810	.819	.835	.695	.723
Unweighted average	MCORES	.859	.837	.875	.832	.897	.639	.750	.748	.732	.767
	Baseline	.688	.832	.851	.823	.820	.600	.763	.793	.750	.741
Recall											
MUC	MCORES	.886	.887	.936	.875	.925	.841	.796	.779	.460	.838
	Baseline	.786	.886	.890	.814	.910	.869	.850	.857	.587	.869
B^3	MCORES	.894	.948	.958	.971	.960	.782	.921	.929	.961	.930
	Baseline	.802	.945	.957	.973	.953	.754	.923	.934	.961	.929
CEAF	MCORES	.753	.948	.963	.979	.950	.838	.944	.945	.956	.546
	Baseline	.764	.946	.953	.975	.946	.827	.945	.951	.964	.535
BLANC	MCORES	.800	.743	.788	.732	.960	.538	.637	.666	.651	.930
	Baseline	.589	.735	.783	.750	.597	.528	.641	.673	.659	.937
Unweighted average	MCORES	.833	.882	.911	.889	.949	.750	.825	.830	.757	.811
	Baseline	.768	.878	.896	.878	.852	.744	.840	.854	.793	.818
F-measure											
MUC	MCORES	.892	.737	.794	.641	.856	.627	.540	.582	.406	.624
	Baseline	.721	.725	.776	.643	.726	.544	.568	.628	.451	.574
B^3	MCORES	.918	.968	.976	.984	.975	.867	.955	.959	.976	.961
	Baseline	.882	.967	.974	.984	.972	.863.	.958	.963	.978	.964
CEAF	MCORES	.707	.917	.937	.975	.913	.388	.871	.888	.951	.581
	Baseline	.404	.913	.932	.975	.862	.340	.876	.896	.956	.564
BLANC	MCORES	.851	.784	.837	.780	.849	.568	.687	.682	.643	.831
	Baseline	.645	.778	.797	.768	.655	.551	.696	.729	.676	.816
Unweighted average	MCORES	.842	.852	.886	.845	.898	.613	.763	.778	.744	.749
	Baseline	.663	.846	.870	.843	.804	.575	.775	.804	.765	.730

Table 8: Precision (P), recall (R), and F-measure (F) evaluation of pair classification for MCORES and the baseline. Evaluated over three types of markables, based on the degree of token overlap: exact overlap, partial overlap, and no overlap

Token Overlap	System	Persons			Problems			Treatments			Tests			Across all markables		
		P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Per-entity Runs																
None	**MCORES**	.857	.848	.852	.045	.844	.086	.027	.750	.053	.000	.000	.000	.732	.848	.786
	Baseline	.322	.674	.435	.000	.000	.000	.000	.000	.000	.000	.000	.000	.273	.674	.388
Partial	**MCORES**	.760	.818	.788	.768	.802	.785	.808	.842	.824	.741	.725	.733	.776	.808	.792
	Baseline	.125	.933	.220	.756	.776	.766	.825	.721	.769	.816	.661	.730	.702	.746	.723
Exact	**MCORES**	.996	.985	.990	1.000	.971	.985	1.000	.974.	.987	.995	.967	.981	.997	.980	.988
	Baseline	1.000	.971	.985	1.000	.971	.985	1.000	.974	.987	1.000	.967	.983	1.000	.971	.985
Per-corpus Runs																
None	**MCORES**	.271	.763	.400	.011	.113	.020	.012	.108	.022	.026	.188	.045	.669	.190	.296
	Baseline	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
Partial	**MCORES**	.683	.714	.698	.445	.774	.565	.595	.756	.666	.559	.686	.616	0.542	0.740	0.626
	Baseline	.077	.932	.142	.183	.713	.291	.166	.671	.267	.124	.621	.206	0.151	0.706	0.249
Exact	**MCORES**	.949	.863	.904	.886	.968	.925	.852	.957.	.901	.901	.743	.814	0.920	0.884	0.901
	Baseline	1.000	.639	.780	1.000	.964	.982	1.000	.950	.974	1.000	.653	.790	1.000	0.719	0.837