ENCORE: Experiments with a Synthetic Entity Co-reference Resolution Tool

Bo Lin, Rushin Shah, Robert Frederking, Anatole Gershman
Language Technologies Institute, School of Computer Science, Carnegie Mellon University
5000 Forbes Ave., Pittsburgh 15213, PA, USA
{bolin, rnshah, ref+, anatoleg}@cs.cmu.edu

Abstract

We present ENCORE, a system for entity co-reference resolution that synthesizes the outputs of several off-the-shelf co-reference resolution systems. To boost precision, we filter the output using a named entity recognition tool called SYNERGY which itself is a synthesis of several off-the-shelf NER systems. ENCORE is designed to work under two conditions: NP-CR which resolves noun phrase co-reference and NE-CR which resolves co-references only for named entities. We report the results of our experiments with ENCORE that show 2% to 40% improvements in precision, recall and F-scores over the underlying systems. This opens a promising approach which leverages the existing "black box" state-of-the-art tools without attempting to re-create their achievements and focuses the development efforts on the differences in their output.

1. Introduction

"Co-reference resolution is the process of determining whether two expressions in natural language refer to the same entity in the world" (Soon et al., 2001). It is a critical task in information extraction and it has received much attention in the last decade using both rule-based and machine learning approaches. As a result, there is a growing number of proprietary and open-source co-reference resolution systems (Versley et al., 2008; Bengston & Roth, 2008). Their performance is typically in the 50% to 70% F-measure range on various metrics, which leaves substantial room for improvement. More importantly, many tools tend to specialize in particular areas such as foreign names or biological entities. It would be virtually impossible to re-create all of their best features in a single tool. A contrasting approach would be to create a system capable of using a variety of available tools as black boxes, leveraging their individual capabilities, integrating and improving their collective results. This approach is challenging because the underlying tools were not created for the purpose of integration. The tools also change as their authors introduce new features and improve performance.

In this paper, we propose a novel synthetic tool called ENCORE that provides superior performance by leveraging several of the best state-of-the-art tools. We treat the underlying tools as co-reference annotators and developed several heuristics that examine their results and create synthetic co-reference classes. Our tests show 2% to 27% and 20% to 40% improvements under two different conditions over the underlying systems in precision, recall and F-scores on two test sets. Section 2 of this paper includes a brief review of related work, Section 3 describes our evaluation metrics, Section 4 introduces our integration methodology, and Section 5 provides the results of our experiments.

Finally, a word on terminology: we use the terms "textual reference," "reference" and "mention" interchangeably to refer to a text phrase. We use the terms "entity," "object" and "equivalence class" to refer to real-world objects.

2. Related Work

Current state-of-the-art approaches include both rule-based and machine learning algorithms. The rule-based approaches apply inductive logic programming, which combines rules for co-reference resolution in a logic induction framework. Other researchers use Markov logic networks with a probabilistic version of logic induction (Culotta et al., 2007). Many researchers have explored machine learning approaches by treating the problem as a pair-wise binary classification problem with subsequent entity clustering or a joint model of classification and clustering (Soon et al., 2001; Ng & Cardie, 2002; Ng, 2005; Haghighi & Klein 2007; Ng, 2008; Finkel & Manning, 2008). The most recent work (Haghighi & Klein, 2009) focuses on feature analysis with a simple model for co-reference resolution. With its rich set of syntactic and semantic features, it is reported to outperform the current state-of-the-art systems.

In the work reported in this paper, we focus on the leveraging of publicly available tools for co-reference resolution. One of them is from a recent study on the value of using rich features for co-reference resolution (Bengtson & Roth, 2008) which comes as a Learning-Based Java (LBJ) co-reference package. Another is called BART, which is from the Johns Hopkins University summer workshop on using lexical and encyclopedic knowledge for entity disambiguation (Versley et al., 2008). We used both of these tools as described in Section 4.

3. Evaluation Metrics

Evaluation for co-reference resolution is challenging, considering that co-reference resolution is neither a traditional classification problem nor a labelling problem. A good evaluation metric has to consider both entity recognition and clustering. Several efforts have been made to establish standard evaluation metrics. Link-based F-measure is the one of the earliest metrics adopted in the MUC task (Vilain et al., 1995). In this metric the F-measure is computed on the co-reference links from the

system output against the links in the gold standard. However, it is reported to be biased for systems with fewer entity outputs (Finkel & Manning, 2008; Luo, 2005). Another metric, called B-cubed (B^3) takes the weighted sum of the F-measures for each individual mention which helps alleviate the bias in the pure link-based F-measure (Bagga & Baldwin, 1998). The ACE named entity detection and tracking task used a metric that normalizes the sum of false-alarm, missed and mistaken entities. However, this metric is believed to be non-intuitive and hard to interpret (Luo, 2005). Instead Luo proposed a new metric named Constrained Entity Alignment F-measure (CEAF), which is claimed to be both discriminative and interpretable. CEAF uses a set similarity measure $\Phi(A, B)$ that is simply the number of common elements in sets *A* and *B*. CEAF compares the equivalence classes *R* produced by the system to the classes in the gold standard *G*. It calculates the mapping $g:R \rightarrow G$ that maximizes the sum of all $\Phi(R_i, g(R_i))$. The optimal mapping is used to define precision, recall and F-measure as follows:

$$p = \frac{\sum \Phi(R_i, g(R_i))}{\sum |R_i|}$$

$$r = \frac{\sum \Phi(R_i, g(R_i))}{\sum |G_i|}$$

$$f = \frac{2pr}{p+r}$$

In this paper, we present our results using two widely used metrics, B^3 and CEAF, and show our improvement over the baselines.

In addition to the selection of the appropriate metrics, we address the issue of the target entities selection for the co-reference resolution task. Typically, the Entity Co-reference Resolution (ECR) systems try to extract all potential references to real-world objects, including both named and nameless objects. For example, the phrase "a Canadian company" is considered a reference, albeit to a nameless object. The documents in the ACE collection are annotated according to this convention. In our conversations with many potential users of the ECR systems we noted that their major interest was in named objects only. For example, a reference to "a Canadian company" would not be interesting unless that company could be identified by name elsewhere the text. Correspondingly, we established two experimental conditions: the traditional condition of NP-CR (Noun Phrase Co-reference Resolution) and NE-CR (Named Entities Co-reference Resolution). For NE-CR, we manually re-annotated the original gold standard to include only NE-CR references.

4. ENCORE Details

In this section, we present the details of our synthetic co-reference resolution system. The system uses several off-the-shelf co-reference resolution tools and integrates their results. We first discuss these state-of-the-art co-reference resolution tools and then propose our own integration methodology.

4.1 State-of-the-art Tools

For our experiments we selected four popular, publicly available ECR packages: (1) the Learning-Based Java ("LBJ System") co-reference package from UIUC (Bengtson & Roth, 2008), (2) BART co-reference toolkit from Johns Hopkins University Summer Workshop (Versley et al., 2008) and two others. In our testing, LBJ and BART produced results uniformly superior to the other two systems with very little new information provided by the latter two. For this reason, we used only LBJ and BART in ENCORE. LBJ and BART embody different approaches to co-reference resolution. While LBJ incorporates a rich set of syntactic and semantic features, BART additionally uses information from Wikipedia for co-reference disambiguation (Versley et al., 2008). Our expectation is that the different approaches adopted by LBJ and BART would produce complementary results with room for further refinement. We conducted a pilot test on a manually annotated corpus consisting of 10 articles from politics, sports and entertainment with about 500 entities and 100 equivalence classes. The pilot showed that the two systems are rather complementary because a simple union of the co-reference chains discovered by each system resulted in a 10% increase in recall. The pilot also gave us ideas on possible heuristics which we will detail in the following sections. On other sets (see below), we achieved an even greater increase in recall. This number indicates an upper bound for the integration heuristics if they can prevent a deterioration of precision caused by false positives.

4.2 Integration Methodology

In this section, we present our initial integration heuristics for combining the outputs of the primary co-reference resolution tools. We treat each primary system as an annotator that marks text phrases with the labels of the corresponding entities (objects). Each entity produced by the annotator has a set of textual references – some that contain names and some that do not.

First, ENCORE tries to merge entities produced by different annotators. The first heuristic is rather obvious: if two such entities have identical lists of textual references, they are merged. The second heuristic is more interesting and important. We merge two entities if their lists of textual references have at least one *pivot* reference in common. Which reference qualifies as a pivot reference is of critical importance. Allowing all references to serve as pivots, often leads to serious mistakes. For example, a common pronoun could merge

two entities which may not be the same, as in:

(John$_1$, he$_2$) and (he$_2$, Bill$_3$)

On the other hand, if the selection criteria for pivot references are too restrictive then we may end up with too many entities and many ambiguous references like "he$_2$" in the above example, attached to two different entities.

ENCORE uses different pivot criteria in our two experimental conditions. Under the NE-CR condition, only named textual references are used as pivots. Our preliminary investigation indicated that the nominal (nameless) textual references where there is no agreement between the primary co-reference resolution tools are most likely to cause errors if we use them as pivots. Their elimination improves precision without significant reduction of recall. To identify named references we use SYNERGY, our own NER (Named Entity Extractor) (Shah et al., 2010) which synthesizes the results from several off-the-shelf NER systems, but any other high-performance NER system could be used as well.

Under the NP-CR condition, the above restriction on pivots turned out to be too narrow, leaving out too many valid references, especially the classes consisting of only the nominal textual references. For this condition we developed a more "relaxed" version, counting as pivots all phrases that either contain a named reference or are contained in one. For example, the phrase "the President of the United States" is not a named reference (the president is not identified) but it contains a named reference "the United States" which makes it a pivot reference. The textual reference "Properties," while not a named reference, can serve as a pivot if it is contained in the phrase "Hong Kong Properties LTD". In our experiments, mentions "associated" with named references in the above manner were more reliable as pivots, increasing the accuracy of merging.

After the mergers, each named reference belongs to one and only one entity. Nameless references might belong to several entities. ENCORE cleans up the results using the following heuristics. Nameless references that belong to more than one entity are eliminated as ambiguous. Under the NE-CR condition, entities (sets of references) that contain no named references are eliminated.

The following example illustrates how ENCORE works in general under both conditions:

LBJ produces the following two equivalence classes (among many others):

(J. Smith, Joe Smith)

(an oil company)

BART system produces only one class containing

(J. Smith, Mr. Smith)

The reference "J. Smith" is recognized by SYNERGY as a named entity, and the first two classes are merged, producing:

(J. Smith, Joe Smith, Mr. Smith)

The reference "an oil company" is not recognized by SYNERGY and is dropped.

The following example illustrates under NE-CR, how ENCORE is able to improve on the underlying systems without getting confused by a significant error in one of them:

> Our friend **Jakaya Kikwete**$_1$ studied at that school last summer.... By the way, **he**$_2$ is marrying **Maria Kashonda**$_3$ soon.... **Jakaya**$_4$ and **Maria**$_5$ are moving to Teheran where **he**$_6$'ll be working for Pishgaman Nano Arya.

LBJ produces three classes where it misclassifies the last "he" as a reference to "Maria" which is clearly a female name.

LBJ1 = (Jakaya Kikwete$_1$, Jakaya$_4$)

LBJ2 = (Maria Kashonda$_3$, Maria$_5$, he$_6$)

LBJ3 = (he$_2$)

BART returns the following classes:

BART1 = (Jakaya Kikwete$_1$, he$_2$, Jakaya$_4$, he$_6$)

BART2 = (Maria$_5$)

ENCORE merges LBJ1 with BART1 and LBJ2 with BART2. It does not merge LBJ2 and BART1 because "he$_6$" is not a pivot reference. It then eliminates LBJ3 and "he$_6$" producing:

ENT1 = (Jakaya Kikwete$_1$, he$_2$, Jakaya$_4$)

ENT2 = (Maria Kashonda$_3$, Maria$_5$)

In the process, we lost the reference he$_6$ which should have been part of ENT1, but was eliminated because of its ambiguity. A more sophisticated heuristic aware of first name genders would have salvaged it.

5. Experiments

We conducted our experiments on two test sets: MIX1 and ACE NWIRE. MIX1 is a small set of 10 articles that reflect one of our target application domains: business news and biographical sketches. This manually annotated set contains approximately 4000 words, 500 entities and 100 co-reference equivalence classes. The ACE NWIRE set from ACE-2 corpus for NIST Automatic Content Extraction program is widely used in co-reference resolution experiments (Mitchell et al., 2003). It is based on 29 articles and contains approximately 20,000 words, 2600 entities and 1000 co-reference equivalence classes. MIX1 set was too small for meaningful tests under the NE-CR condition. To conduct the tests, we manually created the annotations of the ACE NWIRE corpus with equivalence classes consisting of only named entities and their references (both named and nameless).

Some earlier systems (Bengston & Roth, 2008; Haghighi & Klein 2009) achieved good results, but they matched only head nouns between entities in the gold standard and those in the system output, instead of the entire mentions.

We should note that using entire mentions instead of just head nouns makes evaluation much stricter. We follow this approach, for two reasons: Firstly, our main objective is to show the improvement over baseline systems. By placing strict rules of evaluation, it would be more effective to observe the direct improvements from ENCORE. Secondly, the underlying primary systems, as black boxes, return only the full extends of textual mentions/references with no head noun information for our evaluation.

	Precision	Recall	F-Score
ENCORE	**0.468**	**0.462**	**0.458**
Union	0.256	**0.530**	0.340
LBJ	0.342	0.479	**0.396**
BART	0.404	0.400	**0.387**

Table 1: MIX1 with CEAF under NP-CR

	Precision	Recall	F-Score
ENCORE	**0.410**	**0.425**	**0.395**
Union	0.220	**0.548**	0.301
LBJ	0.346	0.377	**0.356**
BART	0.352	0.354	**0.333**

Table 2: MIX1 with B^3 under NP-CR

For comparison, we used three baselines: the individual performances of LBJ and BART, and another baseline named Union created by us. Union measures the results of a simple union of the equivalence classes retrieved by the two underlying systems. Union should give us the maximum recall achievable by the integration of the underlying systems at the cost of diminished precision. The results of our experiments are shown in Table 1 and Table 2 for the MIX1 test set under the NP-CR condition, Table 3 and Table 4 for the ACE NWIRE test set also under the NP-CR condition, and Table 5 and Table 6 for the ACE NWIRE test set under the NE-CR condition.

5.1 Tests under NP-CR Condition

As we can see from these tables, ENCORE produces better precision, better F-score and comparable recall results as compared to the two underlying systems. On the MIX1 test set under NP-CR condition, ENCORE shows 16% and 18% improvement with CEAF over the individual F-scores of LBJ and BART. Similar improvements of 19% and 27% are observed with B^3. However, the recall results of ENCORE are significantly below the maximum, as indicated by the Union baseline. The 0.462 recall value of ENCORE is about 13% below the recall value of the Union baseline with CEAF. With B^3 the difference is 23%. Our current heuristics reject many valid nominal (nameless) mentions. This leaves ample headroom for future improvements with better heuristics and the use of machine learning techniques for synthesis.

	Precision	Recall	F-Score
ENCORE	**0.525**	**0.521**	**0.512**
Union	0.417	**0.546**	0.465
LBJ	0.485	0.505	**0.493**
BART	0.441	0.386	**0.407**

Table 3: ACE NWIRE with CEAF under NP-CR

	Precision	Recall	F-Score
ENCORE	**0.511**	**0.476**	**0.481**
Union	0.393	**0.536**	0.441
LBJ	0.533	0.446	**0.476**
BART	0.391	0.320	**0.340**

Table 4: ACE NWIRE with B^3 under NP-CR

As shown in the F-Score column of Table 3 and 4, ENCORE improves the F-scores on the ACE NWIRE by 1.7% and 24.8% over LBJ and BART respectively with CEAF. It also gives improvements of 2.7% and 29.0% over the two baselines with B^3.

The improvements on the ACE NWIRE test set is considerably lower than that on MIX1 test set. Close examination of the results on the individual files of the ACE NWIRE set reveals that on this test set, LBJ consistently outperforms BART for most of the files. Yet the latter system still contributes enough differences to improve the precision score over the LBJ system by almost 10% from its 0.485 to 0.525 under CEAF metrics, which may be important for some applications. The headroom for improvements in recall is similar to MIX1: 11%-15%.

5.2 Tests under NE-CR Condition

Tests under the NE-CR condition show significantly better improvements: ENCORE improves the F-scores on the ACE NWIRE by 31% and 40% with CEAF and 19% and 41% with B^3 over LBJ and BART respectively. While the comparison with LBJ and BART under the NE-CR condition might not be entirely fair because these systems were not optimized for this condition, we show how the "black box" systems can be successfully re-purposed for a different task.

	Precision	Recall	F-Score
ENCORE	**0.534**	**0.559**	**0.542**
Union	0.251	**0.572**	0.347
LBJ	0.332	0.556	**0.413**
BART	0.340	0.464	**0.388**

Table 5: ACE NWIRE with CEAF under NE-CR

	Precision	Recall	F-Score
ENCORE	**0.495**	**0.483**	**0.476**
Union	0.270	**0.527**	0.349
LBJ	0.369	0.470	**0.401**
BART	0.329	0.369	**0.337**

Table 6: ACE NWIRE with B^3 under NE-CR

6. Discussion and Conclusion

In this paper, we introduced ENCORE - our system for entity co-reference resolution based on the synthesis of the results from off-the-shelf co-reference resolution and named entities extraction products. The preliminary experiments we conducted on two test sets under more standard NP-CR condition show marked improvements in the F-scores ranging from 2% to 27% with significant headroom for further improvement. Under the NE-CR condition, the experiments show even better performance improvements of 20% to 40% in the F-scores over the baselines. Our main contribution is to show how the growing number of "black box" off-the-shelf systems can be leveraged to create fast prototypes with superior performance even for the tasks that deviate from their original purpose. Instead of re-creating the existing methods, we focused our efforts on the analysis of their short-comings. Our first targets were the discrepancies between the underlying primary systems. Our heuristics based on a named entity filter proved to be quite effective. We are currently investigating additional heuristics and machine learning approaches to synthesize the primary systems which would further improve the performance of ENCORE.

7. References

Bagga, A. & Baldwin, B. (1998). Algorithms for Scoring Coreference Chains. In *Proceedings of the International Conference on Language Resources and Evaluation*.

Bengtson, E. & Roth, D. (2008). Understanding the value of features for coreference resolution. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Culotta, A., Wick, M., Hall, R. & McCallum, A. (2007). First-order probabilistic models for coreference resolution. In *Proceedings of the Annual Conference the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*.

Finkel, J.R. & Manning, C.D. (2008). Enforcing Transitivity in Coreference Resolution. In *Proceedings of the Annual Conference the Association for Computational Linguistics - Human Language Technologies*.

Haghighi, A. & Klein, D. (2009). Simple Coreference Resolution with Rich Syntactic and Semantic Features. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Haghighi, A. & Klein, D. (2007). Unsupervised Coreference Resolution in a Nonparametric Bayesian Model. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Luo, X. (2005). On coreference resolution performance metrics. In *Proceedings of the International Conference on Language Resources and Evaluation - Human Language Technologies*.

McCallum, A. & Wellner, B. (2003). Towards conditional models of identity uncertainty with application to proper noun coreference. In *Proceedings of International Joint Conference of Artificial Intelligence Workshop on Info Integration on the Web*.

Mitchell, A., Strassel, S., Przybocki, M., Davis, J., Doddington, G., Grishman, R., Meyers, A., Brunstein, A., Ferro, A., and Sundheim, B. (2003). *ACE-2 Version 1.0*. Linguistic Data Consortium, Philadelphia.

Ng, V. (2005). Machine learning for coreference resolution: From local classification to global ranking. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Ng, V. (2008). Unsupervised Models for Coreference Resolution. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*.

Ng, V. & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Shah, R., Lin, B., Gershman, A. & Frederking, R. (2010). SYNERGY: A Named Entity Recognition System for Resource-scarce Languages such as Swahili using Online Machine Translation. In *Proceedings of International Conference on Language Resource and Evaluation Workshop on African Language Technology*.

Soon, W.M., Ng, H.T., & Lim, D.C.Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, Volume 27, Issue 4 (December 2001).

Versley, Y., Ponzetto, S.P., Poesio, M., Eidelman, V., Jern, A., Smith, J., Yang, X., & Moschitti, A. (2008). BART: A Modular Toolkit for Coreference Resolution. In *Proceedings of the Annual Conference the Association for Computational Linguistics - Human Language Technologies*.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *Proceedings of Message Understanding Conference-6*.