Unsupervised Event Coreference for Abstract Words

Dheeraj Rajagopal and **Eduard Hovy** and **Teruko Mitamura**
Language Technologies Institute
Carnegie Mellon University
dheeraj@cs.cmu.edu, hovy@cmu.edu, teruko@cs.cmu.edu

Abstract

We introduce a novel approach for resolving coreference when the trigger word refers to multiple (sometimes non-contiguous) clauses. Our approach is completely unsupervised, and our experiments show that Neural Network models perform much better (about 20% more accurate) than traditional feature-rich baseline models. We also present a new dataset for Biomedical Language Processing which, with only about 25% of the original corpus vocabulary, still captures the essential distributional semantics of the corpus.

1 Introduction

Event coreference is a key module in many NLP applications, especially those that involve multisentence discourse. Current event coreference systems restrict the problem to finding a correspondence between trigger words or phrases and their fully coreferent event (word or phrase). This approach is rather limited since it does not handle the case when the trigger refers to several events as a group, as in

> We worked hard all our lives. But one year we **went on vacation. There was boating, crazy adventure sports, and pro-golfing. We also spent time in the evenings strolling around the park**. But eventually we had to go home. There couldn't have been a better **vacation**.

In this paper we generalize the idea of coreference to 3 levels based on the degree of abstraction of the coreference trigger:

1. Level 1 – Direct Mention: The trigger phrase is specific and usually matches the referring event(s) word-for-word or phrase-for-phrase.

2. Level 2 – Single Clause: While there is a similar word-to-phrase or word-to-word relationship as in level 1, the trigger is a more generic event compared to level 1.

3. Level 3 – Multiple Clauses: The trigger is quite generic and refers to a particular instance of an event that is described over multiple clauses or sentences (either contiguous or non-contiguous). Typically, the abstract event refers to a set of [sub]events, each of them with its own own participants or arguments.

See Table 1 for examples.

We use PubMed[1] as our primary corpus.

Almost all work on event coreference (for example, (Liu et al., 2014) (Lee et al., 2012)) applies to levels 1 or 2. In this paper, we propose a generalized coreference classification scheme and address the challenges related to resolving level-3 coreferences.

Creating gold-standard training and evaluation materials for such coreferences is an uphill challenge. First, there is a significant annotation overhead and, depending on the nature of the corpus, the annotator might require significant domain knowledge. Each annotation instance might require multiple labels depending the number of abstract events mentioned in the corpus. Second, the vocabulary of the corpus is rather large due to domain-related

[1] http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

Proceedings of EMNLP 2016 Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods, pages 22–26,
Austin, TX, November 5, 2016.

level 1	In turn the activated **ERK phosphorylates Stim1 at serine 575**, and **this phosphorylation** enhances complex formation of Stim1
level 2	(a) **BMI1 enhances hTERT activity**. (b) **This effect** was attenuated by PTEN , PTEN (CS) , PTEN (GE) , and C-PTEN.
level 3	(a) **To determine whether these clumps were also associated with the cell cortex, we used confocal microscopy.** (b) **The actin clumps were found associated with the cell cortex in only a minority of cases (Fig . 4).** (c) Immuno-EM using anti-actin antibodies has verified **this observation**

Table 1: Examples of various levels of Coreference (triggers are underlined and referent indicated in bold)

named entities like proteins, cell-types, DNA and RNA names. The large vocabulary size necessitates longer and sparser vectors for representing the documents, resulting in significant data sparsity. Last, evaluating such a system in an unsupervised setting usually leads to debatable justifications for evaluating the models. We address these challenges in the following ways:

1. We construct a new dataset, derived from the PubMed corpus, by replacing all named entities with their respective NE label. We normalize Proteins, Cell lines, Cell types, RNA and DNA names using the tagger described in (Tsuruoka et al., 2005). Also, we normalize all *figure* and *table* references to "internal_link", and citations to other studies as "external_link". This significantly reduces the vocabulary of the dataset.

2. We present an unsupervised model to represent abstract coreferences in text. We present multiple baseline systems using the traditional *Bag-of-Words* model and a Neural Network architecture that outperforms the baseline models.

3. We define a cloze-test evaluation method that requires no annotation. Our procedure stems from the following insight. Instead of starting with the coreference trigger word/phrase and asking "which clauses can refer to this?", we train an algorithm to *predict* for a given clause which trigger word/phrase it would 'prefer to' link to, and then apply this algorithm to [sequences of] clauses within the likely scope of reference of a trigger. An example is shown in Table 2. A similar idea was mentioned in (Hermann et al., 2015).

Passage : BAF57 has been shown to directly interact with the androgen and estrogen receptors. We used co-immunoprecipitation experiments to test whether BAF57 forms a complex with PR in cultured cells. In the absence of hormone, a certain proportion of BAF57 already coprecipitated with PR probably due to the large proportion of PR molecules already present in the nucleus in the uninduced state; however 30 minutes after hormone addition the extent of coprecipitation was increased. In contrast, no complex of PR with the PBAF specific subunit, BAF180 was observed independently of the addition of the hormone. As a positive control for *ABSTRACT_COREF_EVENT* we used BAF250, a known BAF specific subunit.
Task: Predict *ABSTRACT_COREF_EVENT* from the list of all abstract events of interest
Answer: this experiment

Table 2: A sample cloze-test evaluation task

2 Related Work

Entity coreference has been studied quite extensively. There are primarily two complementary approaches. The first focuses mainly on identifying entity mention clusters (see (Haghighi and Klein, 2009), (Raghunathan et al., 2010), (Ponzetto and Strube, 2006), (Rahman and Ng, 2011), (Ponzetto and Strube, 2006)). These models employ feature-rich approaches to improve the clustering models and are limited to noun pairs. The second focuses on jointly modeling mentions across all the entries in the document (see (Denis et al., 2007), (Poon and Domingos, 2008), (Wick et al., 2008) and (Lee et al., 2011)). Some more recent work uses event argument information to assist entity coreference; this includes (Rahman and Ng, 2011), (Haghighi and

Klein, 2010).

The distinct problem of Event Coreference has been relatively underexplored. Some earlier work in this area includes (Humphreys et al., 1997) but the work was very specific to selected events. More recently, there have been approaches to model event coreferences separately (Liu et al., 2014) as well as jointly with entities (Lee et al., 2012). All this work makes the limiting assumption of word/phrase to word/phrase coreference (levels 1 and 2 described earlier). Our work aligns with the event coreference literature but assumes longer spans of text and tackles the more challenging problem of abstract multi-event/clause coreference.

3 Model

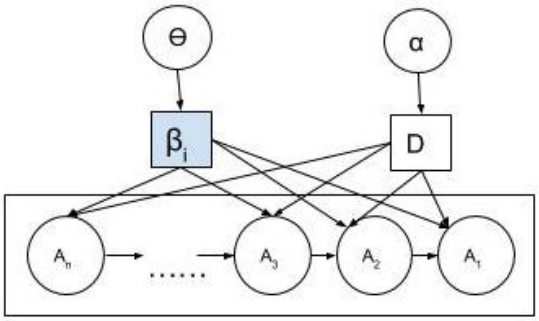

Figure 1: Architecture Diagram for the Coreference Model

Let β_i be the coreference word/phrase generated from a distribution parameterized by θ. Each β_i generates antecedents $A_{1..n}$ (sentences that lead towards the coreference) that contain the coreferent span. These antecendents also obey a dependency relationship between two adjacent sentences in discourse. Since multi-clause coreference shows a distinct effect of recency, we also define a decay function D parameterized by α. The decay function D dictates how the level of association of each antecendent varies over increasing sentence distance.

3.1 Distributed Representation of Sentences

To simplify modeling complexity, we first ensure that all the antecedents are represented by vectors of the same dimension. We use the sentence2vec representation from (Le and Mikolov, 2014) to generate a 300-dimensional continuous distributed representation for each sentence in the PubMed corpus. These vectors are trained using gradient descent, with gradients are obtained though back-propagation. This allows us to reduce the parameters that would have been necessary to model the number of words in each sentence. Table 3 shows some example events and their preferred coreference trigger.

phosphorylation	phophorylation, phosporylation, phoshorylation, dephosphorylation, phosphorylations, Phosphorylation, autophosphorylation, phosphorilation, auto-phosphorylation, phosphorylated
ubiquitination	ubiquitylation, ubiquitinylation, polyubiquitination, poly-ubiquitination, SUMOylation, polyubiquitylation, deubiquitination, sumoylation, autoubiquitination, mono-ubiquitination
concluded	speculated, hypothesized, hypothesised, argued, surmised, conclude, postulated, noticed, noted, postulate

Table 3: Top trigger words (left) under Word2Vec similarity for sample events (right)

3.2 Multilayer Perceptron Model

The MultiLayer Perceptron (MLP) model is given by the function $f : R^D \rightarrow R^L$, where D is the size of input vector x and L is the size of the output vector f(x),

$$f(x) = G\left(b^{(2)} + W^{(2)}\left(s\left(b^{(1)} + W^{(1)}x\right)\right)\right) \quad (1)$$

with bias vectors $b^{(1)}$, $b^{(2)}$; weight matrices $W^{(1)}$, $W^{(2)}$ and activation functions G (softmax) and s (tanh).

For our model, the input dimensions are 300-dimensional sentence vectors. We define 6 classes (6 distinct trigger words) for output. The antecedents are represented using a single vector, composed from the *N* chosen input clauses, where we vary *N* from 1 ato 5. For composition we currently use simple average. We assume no decay currently. The architecture diagram is shown in Figure 2.

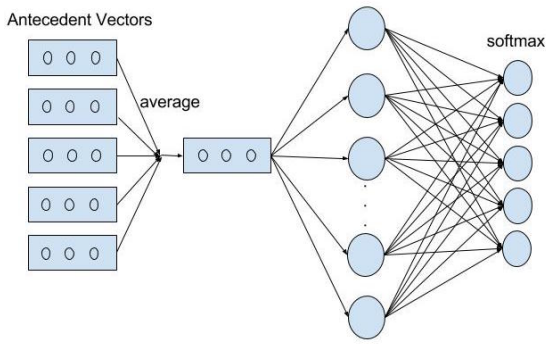

Figure 2: Architecture Diagram for MLP

Classifier	Accuracy
Linear SVM	0.436
SGD Classifer	0.39
BernoulliNB	0.349
Random Forest	0.34
AdaBoost	0.359
DecisionTree	0.286
MLP	**0.62**

Table 4: Results for various baselines and our work

4 Experiments

The Cloze-test evaluation is inspired by the reading comprehension evaluation from Question-Answering research. In this evaluation, the system first reads a passage and then attempts to predict missing words from sentences that contain information from the passage. For our evaluation, we use a slightly modified version of the Cloze-test, in which the model is trained for each coreference with sentences that appear before and after the coreference. Currently, we arbitrarily limit the number of sentences in the antecedent and precedent span for coreference to 5. Also, we consider only 6 labels for now, namely *these changes, these responses, this analysis, this context,this finding, this observation*.

4.1 Experimental Setup

In order to maintain the even distribution of coreference candidates, we derived our dataset from the PubMed corpus by selecting 1000 samples of each of the 6 coreferent labels for a total of 6000 training samples, each sample containing the coreference trigger and we pick antecedent sentences based on the following criteria. If the coreference occurs in the same paragraph, the number of antecedent sentences are limited to sentences from the start of the paragraph or upto five antecedent sentence candidates otherwise. For the MLP model, we use a 70-30 train-test split and apply the early stopping criteria based on accuracy drop on the validation dataset.

4.2 Results

Our results show that our MLP model outperforms all other feature-rich baseline models of traditional classifiers. Although there is general skepticism around sentence vectors, our experiments show that RNN and LSTM models are suitable for the generalized coreference task.

Although we train using a window of N clauses together, during run-time we obtain the prediction for individual sentences rather than taking the average over a window. The label of each sentence or clause depends on the preference of its immediate neighbours, and how these sentences form a 'span', to arrive at a general 'consensus' label. This testing criteria can be further improved by using advanced similarity and coherence detection methods. For now, if the predicted class for that particular sentence is the same as the true label, then that sentence is labeled as part of the coreference.

5 Conclusion and Future Work

We presented a classification taxonomy that generalizes types of event coreference. We presented a model for unsupervised abstract coreference. We described a new dataset for biomedical text that is suitable for any generalized biomedical NLP task. Our Cloze-test evaluation method makes annotation unnecessary.

Since this one of the first works to explore abstract event coreference, there is an uphill task of developing more principled approaches towards modeling and evaluation. We also plan to explore more sophisticated models for our architecture and get more insights into sentence vectors. Also, we plan to extend this idea of coreference into other data domains like News corpus and probably extend to entity-coreference work as well.

Acknowledgments

This research was supported in part by U.S. Army Research Office (ARO) grant W911NF-14-1-0436

under DARPA's Big Mechanisms Program. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of DARPA, ARO, or the U.S. government.

References

Pascal Denis, Jason Baldridge, et al. 2007. Joint determination of anaphoricity and coreference resolution using integer programming. In *HLT-NAACL*, pages 236–243. Citeseer.

Aria Haghighi and Dan Klein. 2009. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1152–1161. Association for Computational Linguistics.

Aria Haghighi and Dan Klein. 2010. Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 385–393. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701.

Kevin Humphreys, Robert Gaizauskas, and Saliha Azzam. 1997. Event coreference for information extraction. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pages 75–81. Association for Computational Linguistics.

Quoc V Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196.

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

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500. Association for Computational Linguistics.

Zhengzhong Liu, Jun Araki, Eduard H Hovy, and Teruko Mitamura. 2014. Supervised within-document event coreference using information propagation. In *LREC*, pages 4539–4544.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 192–199. Association for Computational Linguistics.

Hoifung Poon and Pedro Domingos. 2008. Joint unsupervised coreference resolution with markov logic. In *Proceedings of the conference on empirical methods in natural language processing*, pages 650–659. Association for Computational Linguistics.

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

Altaf Rahman and Vincent Ng. 2011. Coreference resolution with world knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 814–824. Association for Computational Linguistics.

Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Junichi Tsujii. 2005. Developing a robust part-of-speech tagger for biomedical text. In *Panhellenic Conference on Informatics*, pages 382–392. Springer.

Michael L Wick, Khashayar Rohanimanesh, Karl Schultz, and Andrew McCallum. 2008. A unified approach for schema matching, coreference and canonicalization. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 722–730. ACM.