PARMA: A Predicate Argument Aligner

Travis Wolfe, Benjamin Van Durme, Mark Dredze, Nicholas Andrews,
Charley Beller, Chris Callison-Burch, Jay DeYoung, Justin Snyder,
Jonathan Weese, Tan Xu†, and Xuchen Yao
Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, Maryland USA
†University of Maryland, College Park, Maryland USA

Abstract

We introduce PARMA, a system for cross-document, semantic predicate and argument alignment. Our system combines a number of linguistic resources familiar to researchers in areas such as recognizing textual entailment and question answering, integrating them into a simple discriminative model. PARMA achieves state of the art results on an existing and a new dataset. We suggest that previous efforts have focussed on data that is biased and too easy, and we provide a more difficult dataset based on translation data with a low baseline which we beat by 17% F1.

1 Introduction

A key step of the information extraction pipeline is entity disambiguation, in which discovered entities across many sentences and documents must be organized to represent real world entities. The NLP community has a long history of entity disambiguation both within and across documents. While most information extraction work focuses on entities and noun phrases, there have been a few attempts at predicate, or event, disambiguation. Commonly a situational predicate is taken to correspond to either an event or a state, lexically realized in verbs such as "elect" or nominalizations such as "election". Similar to entity coreference resolution, almost all of this work assumes unanchored mentions: predicate argument tuples are grouped together based on coreferent events. The first work on event coreference dates back to Bagga and Baldwin (1999). More recently, this task has been considered by Bejan and Harabagiu (2010) and Lee et al. (2012). As with unanchored entity disambiguation, these methods rely on clustering methods and evaluation metrics.

Another view of predicate disambiguation seeks to link or align predicate argument tuples to an existing anchored resource containing references to events or actions, similar to anchored entity disambiguation (entity linking) (Dredze et al., 2010; Han and Sun, 2011). The most relevant, and perhaps only, work in this area is that of Roth and Frank (2012) who linked predicates across document pairs, measuring the F1 of aligned pairs.

Here we present PARMA, a new system for predicate argument alignment. As opposed to Roth and Frank, PARMA is designed as a a trainable platform for the incorporation of the sort of lexical semantic resources used in the related areas of Recognizing Textual Entailment (RTE) and Question Answering (QA). We demonstrate the effectiveness of this approach by achieving state of the art performance on the data of Roth and Frank despite having little relevant training data. We then show that while the "lemma match" heuristic provides a strong baseline on this data, this appears to be an artifact of their data creation process (which was heavily reliant on word overlap). In response, we evaluate on a new and more challenging dataset for predicate argument alignment derived from multiple translation data. We release PARMA as a new framework for the incorporation and evaluation of new resources for predicate argument alignment.[1]

2 PARMA

PARMA (Predicate ARguMent Aligner) is a pipelined system with a wide variety of features used to align predicates and arguments in two documents. Predicates are represented as mention spans and arguments are represented as coreference chains (sets of mention spans) provided by in-document coreference resolution systems such as included in the Stanford NLP toolkit. Results indicated that the chains are of sufficient quality so as not to limit performance, though future work

[1] https://github.com/hltcoe/parma

RF
• Australian [police]$_1$ have [arrested]$_2$ a man in the western city of Perth over an alleged [plot]$_3$ to [bomb]$_4$ Israeli diplomatic [buildings]$_5$ in the country , police and the suspect s [lawyer]$_6$ [said]$_7$
• Federal [police]$_1$ have [arrested]$_2$ a man over an [alleged]$_5$ [plan]$_3$ to [bomb]$_4$ Israeli diplomatic [posts]$_8$ in Australia , the suspect s [attorney]$_6$ [said]$_7$ Tuesday
LDC MTC
• As I [walked]$_1$ to the [veranda]$_2$ side , I [saw]$_2$ that a [tent]$_3$ is being decorated for [Mahfil-e-Naat]$_4$ -LRB- A [get-together]$_5$ in which the poetic lines in praise of Prophet Mohammad are recited -RRB-
• I [came]$_1$ towards the [balcony]$_2$, and while walking over there I [saw]$_2$ that a [camp]$_3$ was set up outside for the [Naatia]$_4$ [meeting]$_5$.

Figure 1: Example of gold-standard alignment pairs from Roth and Frank's data set and our data set created from the LDC's Multiple Translation Corpora. The RF data set exhibits high lexical overlap, where most of the alignments are between identical words like *police-police* and *said-said*. The LDC MTC was constructed to increase lexical diversity, leading to more challenging alignments like *veranda-balcony* and *tent-camp*

may relax this assumption.

We refer to a predicate or an argument as an "item" with type *predicate* or *argument*. An alignment between two documents is a subset of all pairs of items in either documents with the same type.[2] We call the two documents being aligned the source document S and the target document T. Items are referred to by their index, and $a_{i,j}$ is a binary variable representing an alignment between item i in S and item j in T. A full alignment is an assignment $\vec{a} = \{a_{ij} : i \in N_S, j \in N_T\}$, where N_S and N_T are the set of item indices for S and T respectively.

We train a logistic regression model on example alignmentsand maximize the likelihood of a document alignment under the assumption that the item alignments are independent. Our objective is to maximize the log-likelihood of all $p(S, T)$ with an L1 regularizer (with parameter λ). After learning model parameters w by regularized maximum likelihood on training data, we introducing a threshold τ on alignment probabilities to get a classifier. We perform line search on τ and choose the value that maximizes F1 on dev data. Training was done using the Mallet toolkit (McCallum, 2002).

2.1 Features

The focus of PARMA is the integration of a diverse range of features based on existing lexical semantic resources. We built PARMA on a supervised framework to take advantage of this wide variety of features since they can describe many different correlated aspects of generation. The following features cover the spectrum from high-precision to high-recall. Each feature has access to the proposed argument or predicate spans to be linked and the containing sentences as context. While we use supervised learning, some of the existing datasets for this task are very small. For extra training data, we pool material from different datasets and use the multi-domain split feature space approach to learn dataset specific behaviors (Daumé, 2007).

Features in general are defined over mention spans or head tokens, but we split these features to create separate feature-spaces for predicates and arguments.[3]

For argument coref chains we heuristically choose a canonical mention to represent each chain, and some features only look at this canonical mention. The canonical mention is chosen based on length,[4] information about the head word,[5] and position in the document.[6] In most cases, coref chains that are longer than one are proper nouns and the canonical mention is the first and longest mention (outranking pronominal references and other name shortenings).

PPDB We use lexical features from the Paraphrase Database (PPDB) (Ganitkevitch et al., 2013). PPDB is a large set of paraphrases extracted from bilingual corpora using pivoting techniques. We make use of the English lexical portion which contains over 7 million rules for rewriting terms like "planet" and "earth". PPDB offers a variety of conditional probabilities for each (synchronous context free grammar) rule, which we

[2]Note that type is not the same thing as part of speech: we allow nominal predicates like "death".

[3]While conceptually cleaner, In practice we found this splitting to have not impact on performance.

[4]in tokens, not counting some words like determiners and auxiliary verbs

[5]like its part of speech tag and whether the it was tagged as a named entity

[6]mentions that appear earlier in the document and earlier in a given sentence are given preference

treat as independent experts. For each of these rule probabilities (experts), we find all rules that match the head tokens of a given alignment and have a feature for the max and harmonic mean of the log probabilities of the resulting rule set.

FrameNet FrameNet is a lexical database based on Charles Fillmore's Frame Semantics (Fillmore, 1976; Baker et al., 1998). The database (and the theory) is organized around semantic frames that can be thought of as descriptions of events. Frames crucially include specification of the participants, or Frame Elements, in the event. The Destroying frame, for instance, includes frame elements `Destroyer` or `Cause Undergoer`. Frames are related to other frames through inheritance and perspectivization. For instance the frames `Commerce_buy` and `Commerce_sell` (with respective lexical realizations "buy" and "sell") are both perspectives of `Commerce_goods-transfer` (no lexical realizations) which inherits from `Transfer` (with lexical realization "transfer").

We compute a shortest path between headwords given edges (hypernym, hyponym, perspectivized parent and child) in FrameNet and bucket by distance to get features. We also have a binary feature for whether two tokens evoke the same frame.

TED Alignments Given two predicates or arguments in two sentences, we attempt to align the two sentences they appear in using a Tree Edit Distance (TED) model that aligns two dependency trees, based on the work described by (Yao et al., 2013). We represent a node in a dependency tree with three fields: lemma, POS tag and the type of dependency relation to the node's parent. The TED model aligns one tree with the other using the dynamic programming algorithm of Zhang and Shasha (1989) with three predefined edits: deletion, insertion and substitution, seeking a solution yielding the minimum edit cost. Once we have built a tree alignment, we extract features for 1) whether the heads of the two phrases are aligned and 2) the count of how many tokens are aligned in both trees.

WordNet WordNet (Miller, 1995) is a database of information (synonyms, hypernyms, etc.) pertaining to words and short phrases. For each entry, WordNet provides a set of synonyms, hypernyms, etc. Given two spans, we use WordNet to determine semantic similarity by measuring how many synonym (or other) edges are needed to link two terms. Similar words will have a short distance. For features, we find the shortest path linking the head words of two mentions using synonym, hypernym, hyponym, meronym, and holonym edges and bucket the length.

String Transducer To represent similarity between arguments that are names, we use a stochastic edit distance model. This stochastic string-to-string transducer has latent "edit" and "no edit" regions where the latent regions allow the model to assign high probability to contiguous regions of edits (or no edits), which are typical between variations of person names. In an edit region, parameters govern the relative probability of insertion, deletion, substitution, and copy operations. We use the transducer model of Andrews et al. (2012). Since in-domain name pairs were not available, we picked 10,000 entities at random from Wikipedia to estimate the transducer parameters. The entity labels were used as weak supervision during EM, as in Andrews et al. (2012).

For a pair of mention spans, we compute the conditional log-likelihood of the two mentions going both ways, take the max, and then bucket to get binary features. We duplicate these features with copies that only fire if both mentions are tagged as PER, ORG or LOC.

3 Evaluation

We consider three datasets for evaluating PARMA. For richer annotations that include lemmatizations, part of speech, NER, and in-doc coreference, we pre-processed each of the datasets using tools[7] similar to those used to create the Annotated Gigaword corpus (Napoles et al., 2012).

Extended Event Coreference Bank Based on the dataset of Bejan and Harabagiu (2010), Lee et al. (2012) introduced the Extended Event Coreference Bank (EECB) to evaluate cross-document event coreference. EECB provides document clusters, within which entities and events may corefer. Our task is different from Lee et al. but we can modify the corpus setup to support our task. To produce source and target document pairs, we select the first document within every cluster as the source and each of the remaining documents as target documents (i.e. $N - 1$ pairs for a cluster of size N). This yielded 437 document pairs.

Roth and Frank The only existing dataset for our task is from Roth and Frank (2012) (RF), who

[7] `https://github.com/cnap/anno-pipeline`

annotated documents from the English Gigaword Fifth Edition corpus (Parker et al., 2011). The data was generated by clustering similar news stories from Gigaword using TF-IDF cosine similarity of their headlines. This corpus is small, containing only 10 document pairs in the development set and 60 in the test set. To increase the training size, we train PARMA with 150 randomly selected document pairs from both EECB and MTC, and the entire dev set from Roth and Frank using multi-domain feature splitting. We tuned the threshold τ on the Roth and Frank dev set, but choose the regularizer λ based on a grid search on a 5-fold version of the EECB dataset.

Multiple Translation Corpora We constructed a new predicate argument alignment dataset based on the LDC Multiple Translation Corpora (MTC),[8] which consist of multiple English translations for foreign news articles. Since these multiple translations are semantically equivalent, they provide a good resource for aligned predicate argument pairs. However, finding good pairs is a challenge: we want pairs with significant overlap so that they have predicates and arguments that align, but not documents that are trivial rewrites of each other. Roth and Frank selected document pairs based on clustering, meaning that the pairs had high lexical overlap, often resulting in minimal rewrites of each other. As a result, despite ignoring all context, their baseline method (lemma-alignment) worked quite well.

To create a more challenging dataset, we selected document pairs from the multiple translations that minimize the lexical overlap (in English). Because these are translations, we know that there are equivalent predicates and arguments in each pair, and that any lexical variation preserves meaning. Therefore, we can select pairs with minimal lexical overlap in order to create a system that truly stresses lexically-based alignment systems.

Each document pair has a correspondence between sentences, and we run GIZA++ on these sentences to produce token-level alignments. We take all aligned nouns as arguments and all aligned verbs (excluding be-verbs, light verbs, and reporting verbs) as predicates. We then add negative examples by randomly substituting half of the sentences in one document with sentences from another corpus, guaranteed to be unrelated. The amount of substitutions we perform can vary the "relatedness" of the two documents in terms of the predicates and arguments that they talk about. This reflects our expectation of real world data, where we do not expect perfect overlap in predicates and arguments between a source and target document, as you would in translation data.

Lastly, we prune any document pairs that have more than 80 predicates or arguments or have a Jaccard index on bags of lemmas greater than 0.5, to give us a dataset of 328 document pairs.

[8] LDC2010T10, LDC2010T11, LDC2010T12, LDC2010T14, LDC2010T17, LDC2010T23, LDC2002T01, LDC2003T18, and LDC2005T05

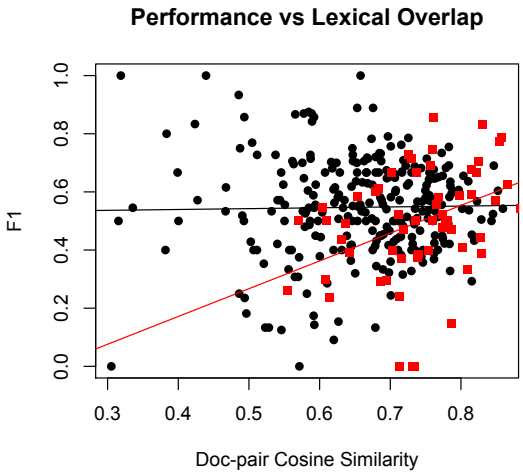

Figure 2: We plotted the PARMA's performance on each of the document pairs. Red squares show the F1 for individual document pairs drawn from Roth and Frank's data set, and black circles show F1 for our Multiple Translation Corpora test set. The x-axis represents the cosine similarity between the document pairs. On the RF data set, performance is correlated with lexical similarity. On our more lexically diverse set, this is not the case. This could be due to the fact that some of the documents in the RF sets are minor re-writes of the same newswire story, making them easy to align.

Metric We use precision, recall, and F1. For the RF dataset, we follow Roth and Frank (2012) and Cohn et al. (2008) and evaluate on a version of F1 that considers SURE and POSSIBLE links, which are available in the RF data. Given an alignment to be scored A and a reference alignment B which contains SURE and POSSIBLE links, B_s and B_p respectively, precision and recall are:

$$P = \frac{|A \cap B_p|}{|A|} \qquad R = \frac{|A \cap B_s|}{|B_s|} \tag{1}$$

		F1	P	R
EECB	lemma	63.5	84.8	50.8
	PARMA	**74.3**	80.5	69.0
RF	lemma	48.3	40.3	60.3
	Roth and Frank	54.8	59.7	50.7
	PARMA	**57.6**	52.4	64.0
MTC	lemma	42.1	51.3	35.7
	PARMA	**59.2**	73.4	49.6

Table 1: PARMA outperforms the baseline lemma matching system on the three test sets, drawn from the Extended Event Coreference Bank, Roth and Frank's data, and our set created from the Multiple Translation Corpora. PARMA achieves a higher F1 and recall score than Roth and Frank's reported result.

and F1 as the harmonic mean of the two. Results for EECB and MTC reflect 5-fold cross validation, and RF uses the given dev/test split.

Lemma baseline Following Roth and Frank we include a lemma baseline, in which two predicates or arguments align if they have the same lemma.[9]

4 Results

On every dataset PARMA significantly improves over the lemma baselines (Table 1). On RF, compared to Roth and Frank, the best published method for this task, we also improve, making PARMA the state of the art system for this task. Furthermore, we expect that the smallest improvements over Roth and Frank would be on RF, since there is little training data. We also note that compared to Roth and Frank we obtain much higher recall but lower precision.

We also observe that MTC was more challenging than the other datasets, with a lower lemma baseline[10]. Figure 2 shows the correlation between document similarity and document F1 score for RF and MTC. While for RF these two measures are correlated, they are uncorrelated for MTC. Additionally, there is more data in the MTC dataset which has low cosine similarity than in RF.

5 Conclusion

PARMA achieves state of the art performance on three datasets for predicate argument alignment. It builds on the development of lexical semantic resources and provides a platform for learning to utilize these resources. Additionally, we show that task difficulty can be strongly tied to lexical similarity if the evaluation dataset is not chosen carefully, and this provides an artificially high baseline in previous work. PARMA is robust to drops in lexical similarity and shows large improvements in those cases. PARMA will serve as a useful benchmark in determining the value of more sophisticated models of predicate-argument alignment, which we aim to address in future work.

While our system is fully supervised, and thus dependent on manually annotated examples, we observed here that this requirement may be relatively modest, especially for in-domain data.

[9]We could not reproduce lemma from Roth and Frank (shown in Table 1) due to a difference in lemmatizers. We obtained 55.4; better than their system but worse than PARMA.

[10]Recall our lemma baseline for RF was 55.4.

Acknowledgements

We thank JHU HLTCOE for hosting the winter MiniSCALE workshop that led to this collaborative work and Vulcan Inc. for funding. This material is based on research sponsored by the NSF under grant IIS-1249516 and DARPA under agreement number FA8750-13-2-0017 (the DEFT program). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of NSF, DARPA, or the U.S. Government.

References

Nicholas Andrews, Jason Eisner, and Mark Dredze. 2012. Name phylogeny: A generative model of string variation. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Amit Bagga and Breck Baldwin. 1999. Cross-document event coreference: Annotations, experiments, and observations. In *Proceedings of the Workshop on Coreference and its Applications*, pages 1–8. Association for Computational Linguistics.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Cosmin Adrian Bejan and Sanda Harabagiu. 2010. Unsupervised event coreference resolution with rich linguistic features. In *Proceedings of the 48th Annual Meeting of the Association for Computational*

Linguistics, ACL '10, pages 1412–1422, Stroudsburg, PA, USA. Association for Computational Linguistics.

Trevor Cohn, Chris Callison-Burch, and Mirella Lapata. 2008. Constructing corpora for the development and evaluation of paraphrase systems. *Comput. Linguist.*, 34(4):597–614, December.

Hal Daumé. 2007. Frustratingly easy domain adaptation. In *Annual meeting-association for computational linguistics*, volume 45, page 256.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Conference on Computational Linguistics (Coling)*.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280(1):20–32.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Xianpei Han and Le Sun. 2011. A generative entity-mention model for linking entities with knowledge base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 945–954. 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 Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/ mccallum/mallet.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *AKBC-WEKEX Workshop at NAACL 2012*, June.

Robert Parker, David Graff, Jumbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English gigaword fifth edition.

Michael Roth and Anette Frank. 2012. Aligning predicate argument structures in monolingual comparable texts: A new corpus for a new task. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 218–227, Montréal, Canada, 7-8 June. Association for Computational Linguistics.

Xuchen Yao, Benjamin Van Durme, Peter Clark, and Chris Callison-Burch. 2013. Answer extraction as sequence tagging with tree edit distance. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

K. Zhang and D. Shasha. 1989. Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, 18(6):1245–1262, December.